Information Retrieval, Text Composition, and Semantics

Birger Hjørland

Royal School of Library and Information Science, Copenhagen

Birger Hjørland is Head of Department of Science and Humanities Information Studies, Royal School of Library and Information Science in Copenhagen. He earned his Ph.d. from the University of Göthenborg in 1993 in Information Science and an MA in psychology from the University of Copenhagen in 1974. He has been a research librarian at the Royal Library in Copenhagen 1978-1990 and the co-ordinator for the library's computer based reference services.

Hjørland, B. (1998). **Information Retrieval, Text Composition, and Semantics.** *Knowledge Organization*, *25(1/2)*, 16-31. 63 refs.[1]

ABSTRACT: Information science (IS) is concerned with the searching and retrieval of text and other information (IR), mostly in electronic databases and on the Internet. Such databases contain fulltext (or other kinds of documents, e.g. pictures) and/or document representations and/or different kinds of "value added information". The core theoretical problem for IS is related to the determination of the usefulness of different "subject access points" in electronic databases. This problem is again related to theories of meaning and semantics.[2] This paper outlines some important principles in the design of documents done in the field of "composition studies". It maps the possible subject access points and presents research done on each kind of these. It shows how theories of IR must build on or relate to different theories of concepts and meaning. It discusses two contrasting theories of semantics worked out by Ludwig Wittgenstein: "the picture theory" and "the theory of language games" and demonstrates the different consequences for such theories for IR. Finally, the implications for information professionals are discussed.

1. Introduction

Information retrieval (IR) is the process in which users put questions to information systems and consequently get some answers (see the model in Ingwersen, 1992, p. 55). At the most elementary level, this interaction consists of 1) a query 2) some text representations 3) some matching technique. The scientific/empirical investigation of IR started about 1950. It has comprised both the processes in computers, and in the users ("the physical paradigm" and "the cognitive paradigm" as Ellis, 1996, names them). What direction should this research take after nearly 50 years of rather intensive research?

In my opinion different views on IR and IS imply different views on cognition, on concepts, and on meaning. It can be difficult to describe the cognitive or the semantic presumptions behind the physical and the cognitive paradigms, respectively. But all techniques and all theories build on some metatheoretical and epistemological assumptions. In IS it has become very important to study the assumptions and implicit theories, with which researchers look at computers, texts, users, questions and interactions. The breakthrough of an important "non rationalistic" or non-positivist interdisciplinary viewpoint was Winograd & Flores (1986). Since then, IS has opened up for many new important and related metatheoretical views (e.g., hermeneutics, phenomenology, social constructivism, semiotics, and activity theory).

Very central in this reorientation in IS are in my opinion both a new focus on meaning and a new focus on the social environments of both users and systems. Van Rijsbergen (1986, p. 194) has pointed out that the concept of meaning has been overlooked in IS, why the whole area is in a crisis. The fundamental basis of all the previous work – including his own – is wrong because it has been based on the assumption that a formal notion of meaning is not required to solve the IR problems. This statement alone should justify a closer cooperation between IS and the multidisciplinary research done in semantics. Leading information scientists have treated semantic problems earlier (e.g., Blair, 1990, Dahlberg, 1978 & 1995, Foskett, 1977, and Vickery & Vickery, 1987), but they have seldom related their research to the theories developed in semantics.

2. Subject Access Points

It is a trivial statement that the IR mechanism must match the query with some specific elements in the documents/texts or their representations in the information systems. However, almost none research has been done to illuminate what kind of documents are produced, and what specific demands such different kinds of documents make to IR systems. It should be a clear goal for IS to make a comprehensive theory of documents, their functions, kinds, structure, etc. In order to simplify things I shall limit myself to one kind of documents: the typical scientific research article.

Table 1

Structure and Elements in a Typical Scientific Article[3]

Norms of scientific method and *Philosophy of science external to the article*	*Elements contained in the article*	*Value added information* *(Subject access points, access and evaluation information)*
	Bibliographical identification (journal name, volume, pages)	Bibliographical description
	Titel	Relations to other editions
	Author(s) with corporate affiliation and address	Identifier
Observation and description	Author abstract	Biographical information
	(Author keywords)	Institutional information
Problem statement	Introduction	Indexer abstracts
	Apparatus and materials	Indexer descriptors
Hypothesis	Method	Classification codes
Experiment	Results	Language codes
Theory building	Discussion	Document type codes
	Conclusion	Editorial comments Links to citing papers, reviews, and criticism
	(Acknowledgements)	"Key word plus", "research fronts"
	References	Information about availability of document
		Evaluation
		Target group

We may imagine a database on the Internet comprising the fulltext editions of all the scientific journals indexed in such databases as Chemical Abstracts, MEDLINE, PsycINFO and SciSearch. In addition to the scientific journals themselves, we have the "value added" information produced by information specialists, by publishers, and by other professionals. Of course, the future publishing of online documents rather than printed documents is going to change both the process of writing ("scholarly skywriting") and the character of the written texts themselves (see, e.g., Harnad, 1990 & 1991). However, as our point of departure we will look at the written texts, as we know them today. An outline of all this information is given in table 1.

Given all this information in an online system we may now look at the system from the searchers' point of view: all the elements in the records are potential "subject access points". If a user is interested in some eating disorder, he or she can choose one database or another, she can search, for example, words in titles, words in abstracts, descriptors, or classification codes in PsycINFO or MEDLINE, search cited references, "key words plus" or "research fronts" in SciSearch, search in all the elements in fulltext databases, and so on. IR is essentially a theory about the most rational and efficient way to design search profiles (or rather "search interactions") and consequently to provide principles on how to organise knowledge in order to maximise its retrievability.

Real IR usually employs combinations of sets of terms. E.g.: "Treating young anorexian females with cognitive therapy" combining "anorexia" and "human females" and ("cognitive therapy" or "behavioural therapy"). However, a combined search can be no more efficient that each of the sets allows. It is very important that each set is clearly defined. The most basic problem in IR is thus related to the informational value of the different access points in the search process. Again, we can simplify and limit ourselves to regarding only one search term in different access points. Table 2 is an example showing the results from a search in PsycINFO done in 1997.

Table 2:

Distribution of references described by the same term in different subject access points

S1	2271	ANOREXIA/TI	[word in document title]
S2	2639	ANOREXIA/ID	[word in identifier]
S3	2963	ANOREXIA/DE	[word in descriptor]
S4	3386	ANOREXIA/AB	[word in abstracts]
S5	4177	S1 OR S2 OR S3 OR S4	[union of sets]
S6	4177	ANOREXIA	[default access=S5]
S0	1508	S1 AND S2 AND S3 AND S4	[intersection of sets]

What kinds of theories exist in the literature of IS concerning the different meanings of such different

fields or access points? My claim is that no such theories exist. Many information scientist have traditionally been more like engineers, seeking solutions like "technical fixes", rather than being philosophers seeking theoretical understanding of underlying phenomena. However, experienced searchers do have a lot of tacit knowledge, which, however, is often limited to particular databases. Further it is my assumption that mainstream IR is influenced by some implicit assumptions closely related to those of logical positivism. My suggestion is therefore to continue the work done by Blair (1990) and others, and try to relate the problems of IR to semantic theories.

3. The Picture Theory of Meaning And Its Relation to Theoretical Assumptions in IR

Things are often most clear and understandable if you can illuminate the problem by means of contrasting theories. Even if things are not that simple, sharp opposition can inspire further research which can lead to more varied theories. Such contrasting theories can be found within the works of the same person: The philosopher Ludwig Wittgenstein (1889-1951). As a young man he had an important influence on the Vienna Circle, which was the mainspring of Logical Positivism.[4] In 1921 he published Tractatus Logico-Philosophicus, containing a semantic theory named "the picture theory". Between 1929 and 1932 his ideas underwent dramatic change, which he consolidated over the next fifteen years. These ideas were given definitive expression in Philosophical Investigations (1953), published two years after his death. The new semantic theory ("the later Wittgenstein") could be labelled "theory of language games". While the early Wittgenstein was connected to the empiricist/ positivist positions in philosophy, the later Wittgenstein is related to ordinary language philosophy and pragmatism. Below are listed some principles of the picture theory, which should give enough impression of its essence:

Some Basic Characteristics of "The Picture Theory"

- The ultimate elements of language are names that designate simple objects.
- The meaning of a word is the thing it stands for.
- The substance of all possible worlds consists of the totality of eternal or sempiternal simple objects such as spatio temporal points, un-analysable properties, and relations.
- The meaning of words in public language derive from the ideas or mental images that words are used to express. The key thing in meaning is the propositional content of the belief or thought that a sentence expresses; this is not essentially derived from communication intentions or from social practices.
- A sentence or proposition is a picture of a (possible) state of affairs; terms correspond to non-linguistic elements, and those terms' arrangements in sentences have the same form as the arrangements of the states of affairs the sentences stand for.
- Descriptive language is the model of language proper.
- Words are – or need to be – sharply defined, analysable by specification of necessary and sufficient conditions of application. Vagueness is regarded as a defect, and there exist absolute standards of exactness.
- All that can be expressed at all, can be said clearly and must have one and only one definite meaning. There are no vague, ambiguous, many valued, implicit or tacit meanings.
- All meaningful sentences are truth functions and extensional. Elementary propositions are the only sentences, which are not truth functions of other sentences. Such elementary sentences are pictures of atomic facts.
- Elementary propositions can be combined to form molecular propositions by means of truth-functional operators—the logical connectives.
- There is an absolute distinction between the simple and the complex.
- The only meaningful sentences are those of (natural) science
- All metaphysical statements are meaningless – including the whole of the tractatus itself! At the same time Tractatus in the preface states that it has basically solved the problems of philosophy!

"The Picture Theory" and related theories have, in my opinion, some very clear and pragmatic consequences for IR. It should be said, however, that this is my interpretation, and that further epistemological studies may be needed. The place here does not allow a detailed discussion. The difficulties in providing such interpretations can be illuminated by pointing out that Wittgenstein himself gave up exemplifying the central concepts and theses in Tractatus. However, in my view it can be argued that the picture theory implies the following principles for IR:

- The meaning of a search term is the same irrespective of the field, in which it is represented. (Principle of semantic atomism #1).
- The meaning of a search term is the same irrespective of its place and context within one document or document representation. (Principle of semantic atomism #2).
- The meaning of a search term is the same irrespective of its scientific domain/discourse, the specific

subject database in which it is represented and other contexts. (Principle of semantic atomism #3).
- Subject analysis is essentially a descriptive process (as opposed to a choice, a decision or an evaluation).
- The more limited a field, the greater is the informational value of a term in that field. (Principle of "semantic condensation").
- The more fields a term is represented in the more relevant is the document, in which the term is represented. (Additive principle #1).
- The more times a term is represented in a given field (e.g. a fulltext field), the greater the likelihood that the document is relevant. (Additive principle #2).
- IR is essentially a question of quantitative/ statistical relationships between sets of terms, which can be executed by computers using algorithmic principles.
- IR is a neutral or value free activity. There are objective, measurable criteria of efficiency/success. (E.g. "recall" and "precision").
- Recall can be improved by having as many different subject descriptions as possible put into the document representations ("the strategy of unlimited aliasing"; see also Brooks, 1993, and Blair, 1990).
- Precision can be improved by using narrower terms, by limiting the search to condensed fields or by combining sets with the logical operators "AND" and "NOT".

Based on these principles the general heuristic lesson from table 2 is that you can increase recall by moving down among these possibilities (S0-S6), and you can increase precision by moving up among them (S6-S0). Such heuristics are not, however, without problems. Examples with other terms provide different results and imply different heuristic rules. Other words have different meanings and can have different distributions. The differences are, for example, much more important and exaggerated if we search the word "female":

S7	128336	FEMALE?
S8	10800	FEMALE?/TI
S9	23483	FEMALE?/DE
S10	73029	FEMALE?/ID
S11	87693	FEMALE?/AB

Female has another distribution because sex is a formal research variable often mentioned in abstracts and identifiers, even if this question is not the central issue in other respects. It is important to know the conventions used by the people producing the respective fields. For example, methods and experimental variables are often mentioned in the ID field, but not as often in the title. When a term, for example, "burnout" is not official, but a kind of slang, it is often used in titles, but never in descriptors (the adequate descriptor in this database is "occupational stress"):

S12	1148	BURNOUT/TI
S13	1261	BURNOUT/ID
S14	0	BURNOUT/DE
S15	996	BURNOUT/AB

Trained human searchers can interpret meanings in search terms and use them in IR in ways which algorithms cannot. Information retrieval has to develop a theory that takes content, meaning, and semantics into account. The example shows that universal quantitative relations among kinds of terms or codes not are sufficient. It is not just a question of getting more or less, but what kinds of studies are selected.

I do not claim that the above mentioned principles derived from a positivistic semantics are simply wrong. On the contrary, all experienced searchers, including myself, are using many of them all the time. However, as the search examples show such a theory cannot account for different examples. What I do claim is that IS needs to consider the limitations of this theory: That an understanding of the limits of a semantic theory like "the picture theory" will enable us to build even more advanced information systems (and do better searches in the existing ones). What we need is a semantic theory, which can guide the development of more effective heuristic rules in IR.

4. Other Theories of Semantics

Theories of semantics can be 1) objectivist (i.e. oriented towards objects, the references of the words) or 2) subjectivist (oriented towards the minds, ideas or concepts of individuals) or 3) oriented towards people's social activities. Socially oriented semantic theories can again be more subjectivist (as social constructivism) or more objectivistic/realistic (as, e.g., scientific realism and activity theory).

The picture theory is very objectivistic when it defines "the meaning of a word is the thing it stands for". However, this can be combined with the view that each individual person forms his or her individual concepts of things in the world, which imply a very subjectivist view of meaning. Such subjectivism (and the mixture of the metaphysics of logical positivism and subjectivism) has had a very strong influence in many sciences, including IS. Woodfield (1991) writes that many theorists in cognitive science assume that the individual subject forms standing conceptions of things. They take a conception of a category to be a file, or package, of information stored in long-term memory. This notion of a conception bears a family resemblance to the ordinary notion, but different from it in significant ways. The case for believing in

such file-like structures is, according to Woodfield, not very strong. An alternative proposal is sketched according to which the subject's conceptions are transient, purpose-relative perspectives on things.

A Simple Classification of Semantic Theories

	Individualistic theories	Socially oriented theories
Subjectivist or mentalistic theories	Meanings are individual constructions. E.g., John Locke, theories about "inner language" or "private language", cognitive theories from Jean Piaget to "cognitive science", and G. Lakoff (1987)	Meanings are social constructions. E.g., "social constructivism".
Objectivistic theories	Meanings are the referents of words, or pictures of a given reality. E.g., the picture theory.	Meanings are human discoveries stabilised in language and culture. E.g., pragmatism, scientific realism, "theory of language games", and activity theory.

Stamper (1987), a database semantics, provides a critique of the mixture between positivism and subjectivism in relation to a standardisation program:

> "The errors in the ANSI-SPARC way of treating semantics are twofold. The basic one is their invocation of naïve metaphysics by their use of the term "conceptual". This belongs to a stance of psychologism, which treats semantics as an investigation of relationships of reference between linguistic expressions and concepts in the minds of people, these concepts being their meanings [note 2 excluded]. A more mysterious and unsatisfactory way of establishing meanings could not be chosen for a scientific treatment of the subject. In addition, despite the totally subjective nature of concepts when you come to investigate them, the same database community assumed that a single conceptual schema sufficed to unite the diverse external schemas of various user groups. Users could employ their own local language by adopting synonyms for items in the conceptual schema, and they could limit their domain of discourse to a subset of the conceptual schema, but they had to accept its overarching structure. Hence we see, despite all the subjectivist language of "concepts", that they also adopt a naïve assumption of a single valid view of the world, a kind of sidelong view of the logical positivists' picture of reality. These two errors reflect the metaphysical assumptions widespread among a scientific community reared on a diet of natural science, engineering and mathematics, where a single objective reality is taken for as granted as readily as a belief in the reality of mathematical concepts. (Stamper, 1987, p. 49).

One leading textbook of IR (Ellis, 1996) has described two main approaches in IR: The archetypal (or physical) and the cognitive. In my interpretation, the archetypal approach tends to be very objectivistic, whereas the cognitive has often been very subjectivist. Ellis writes:

> "The archetypal approach tends to focus on the artefacts or surface representations of knowledge recorded in physical media while the cognitive approach displays the opposite tendency and focuses on the people and on modelling personal knowledge structures. Thus, the insistence in the archetypal approach that the framework of understanding be quantitative and removed from the subjectivity of individual cognition, enabled the approach to deal more thoroughly, and with relative conceptual homogeneity, with the artefacts or representations but at the cost of not being able to engage with problems raised by human cognition and knowledge representation—which are fundamental to the retrieval interaction" (Ellis, 1996, p. 191)

In my view, the fundamental problem for IR (as well as for IS as a whole and for many other disciplines) is that it has been caught between individual objectivism and subjectivism. (Subjectivism corresponding to what Frohmann, 1990, criticises as "mentalism" in IR). A third approach seems mandatory. One such theory is the above-mentioned theory developed by the older Wittgenstein.

Some Basic Characteristics of "The Theory of Language Games"

- Language is not strictly held together by logical structure, but consists of a multiplicity of simpler substructures or language games.
- Sentences cannot be taken as logical pictures of facts and the simple components of sentences do not all function as names of simple objects.
- The words "simple" and "complex" have no absolute meaning: What is simple in one language game can be complex in another.
- There are many different languages with many different structures that could meet quite different specific needs.
- There are countless different uses of what we call "symbols", "words", and "sentences". These different functions should be uncovered by philosophy in order to dissolve metaphysical puzzles.

- Common philosophical views about meaning, about logical atomism, about concepts, about rule following are all the product of a wrong view of language.
- Words do not denote sharply circumscribed concepts, but are meant to mark family resemblance between the objects labelled by the concept.
- Words in our language have only meaning insofar as public criteria for their application exist. Consequently, there can be no inner or private language.
- A language is something you learn, and in learning a language, one is initiated into a form of life.
- The world ultimately determines what language games can be played. (A naturalistic, not a relativistic view)
- Meanings are developed in the use of the words or terms. (A use oriented theory of semantics)
- There is no universal scheme of categories to be unveiled, let alone to be established by a theory. Categories cannot have the absolute universality claimed in theories from Aristotle to Russell's logical types.
- Philosophical clarity can be achieved only piecemeal, context by context; there is no short cut via an ideal language, classification or categorisation. [Implicating a domain oriented approach]
- "The ostensive definition explains the use – the meaning – of the word only when the overall role of the word in the language is clear. Thus, if I know that someone means to explain a colour word to me, the ostensive definition "that is called "sepia"" will help me to understand the word. Only if I know what a colour is, am I fully ready for the meaning of "sepia". Here again, knowing what a colour is means being able to do something, knowing how colour terms are used" (Wittgenstein, 1958, §30).

Wittgenstein's general influence has been enormous. Also in IS there seems to be a growing recognition of his importance. He has been cited 67 times in the library & information science journals indexed by the Social Sciences Citation Index (as of January 1998). These citations include Brier (1996), Frohmann (1990), Karamiiftiioglu (1996, 1997), Tuominen (1997), and Warner (1990). Before discussing the implications of this theory, we shall put it in a somewhat broader perspective.

Forrester (1996, p. 28) describes two major approaches to the psychology of language: *The cognition dominant view* following Descartes, Kant, and mainstream cognitive psychology. According to this view concepts and meaning are produced in the individual minds, "in the head". The information from the senses is shaped according to the architecture of our cognitive apparatus in the brain, and this shaping provides the basis of the meaning of words. *The language dominant view* follows the older Wittgenstein, social constructivism [and, e.g. activity theory]. According to this view concepts and meaning are produced by our social practices. A consequence of social practice is the development of communication, of verbal and non-verbal behaviour, and of concepts. Meanings are first produced "outside the head" and are then, through language, transferred into the individual minds. From this perspective, the central question about semantics is not foremost related to individual objects or to individual minds, but to cultures, to subcultures, to the social division of labour, to discourse communities, to scientific disciplines, and so on. The most adequate theories about semantics therefore seem to be sociocognitive and sociolinguistic rather than just cognitive and linguistic.

In this broader context, the pragmatic traditions in semantics have predated the theory of language games. *Charles Sanders Peirce* (1839-1914) found that the pragmatic meaning theory is "futuristic", interpreting meaning from the point of view of how the determining of meaning can contribute to the fulfilment of goals. He wrote:

> The rational meaning of every proposition lies in the future. How so? The meaning of a proposition [its logical interpretant] is itself a proposition. Indeed, it is no other than the very proposition of which it is the meaning: it is a translation of it. But of the myriads of forms into which a proposition may be translated, what is that one which is to be called its very meaning? It is, according to the pragmaticist, that form in which the proposition becomes applicable to human conduct, ... that form which is most directly applicable to self-control under every situation and to every purpose. This is why he locates the meaning in future time; for future conduct is the only conduct that is subject to self control. (Peirce, 1905).

John Dewey (1859-1952) also wrote about the development of meanings (e.g., 1925, 1939, and 1946). He related theories of semantics to the classical epistemologies:

> ... it should be noted that traditional empiricism has also misread the significance of conceptions or general ideas. It has steadily opposed the doctrine of their *a priori* character; it has connected them with experiences of the actual world. But even more obviously than the rationalism it has opposed, empiricism has connected the origin, content and measure of validity of general ideas with *antecedent* existence. According to it, concepts are formed by comparing particular objects, already perceived, with one another, and

> then elimination the elements in which they disagree and retaining that which they have in common. Concepts are thus simply memoranda of identical features in objects already perceived; they are conveniences, bunching together a varity of things scattered about in concrete experience. But they have to be *proved* by agreement with the material of particular *antecedent* experiences; their value and function are essentially retrospective. Such ideas are dead, incapable of performing a regulative office in new situations. They are "empirical" in the sense in which the term is opposed to scientific – that is, they are mere summaries of results obtained under more or less accidental circumstances. (John Dewey, 1939, p. 883)

For John Dewey languages are only one medium of the communication of meaning. Non-verbal communication, art, and objects are all expressive; they carry meaning, and can be regarded as a kind of language. Each art has its own medium and that medium is especially fitted for one kind of communication. The needs of daily life have given superior practical importance to one mode of communication, that of speech. Different human cultures and needs develop special media to communicate meanings. To me, this view seems closely related to Wittgenstein's theory of "language games", which it predated.

John Dewey not only predated the theory of language games. According to Hardwick (1971) he also had a sharper understanding of how meaning develops in use, and the historical character of this development:

> In this chapter, then, I shall be dealing with Wittgenstein's use of "use". The main task one faces in interpreting Wittgenstein's remarks is understanding clearly what he means in saying that the meaning of a word is its "USE in the language". "Use" suggests activity. I should like to consider, therefore, what it means to say that language is an activity. In doing so, I shall compare Wittgenstein's remarks about language as an activity with the pragmatic conception of language developed by George Herbert Mead and John Dewey. We find in Mead, for example, the idea that language is rooted in gesture; that meaning arises out of social activity. Dewey considers language as an instrument, and words and concepts as tools; the importance of language is in what we can do with it. Both of these approaches are similar to the doctrines of Wittgenstein. Second, I want to show that a more carefully worked out pragmatic conception of language stresses a point which Wittgenstein seems to overlook; namely, that the definition of meaning in terms of activity leads to the notion that the meaning of a word is NOT equated to its use in a particular situation. (Hardwick, 1971, 34-35)
>
> The mere use of a word, in the sense that Wittgenstein deals with it in his examples, ignores the larger fact that the word has a history of meaning. And therefore it would be premature to equate the meaning of a word with its use here now. (Hardwick, 1971, 42-43)

The pragmatic theory of meaning is also developed in *"the cultural historical school in psychology"* also known as *"activity theory"* and *"the sociocognitive view"*. Both pragmatism and activity theory are oriented toward the future, toward the fulfilment of human goals, but activity theory is often more explicit about the fact that different groups of people may have goals which are not in harmony. It also stresses the fact that the developments of meanings are often tied to the development of the means of production, to the social division of labour, and to economic influences. Society consists of many "discourse communities", which develop their own special languages, their kinds of documents, their information systems, their institutions, and their professional roles to maintain their functions.

The production of knowledge, the design of documents, the sublanguages, the databases, the use, the collecting and disseminating of knowledge are always done by specific persons, possessing certain views or theories of knowledge. Such fundamental views of knowledge can be more or less conscious or unconscious. Most often they are unconscious. They develop historically and most often interdisciplinary. The explicit analysis of such theories of knowledge is done in philosophy, in science studies, and in the more theoretical parts of the sciences themselves. In my opinion, such analysis should also be done in IS, because theories of knowledge affect every part of the design and use of information systems. Knowledge of this kind may be the only kind of knowledge in IS which can be generalised from each subject area.

Activity theory is much related to social constructivism because both theories are interested in unravelling how meanings have developed historically. However, pragmatism and activity theory are more "realistic" because they find that some interpretations and classifications are simply more optimal than others for given purposes. The objectivity of knowledge is partly a question of what kind of goals the agents are trying to fulfil (implicitly or explicitly, consciously or unconsciously).

From the point of view of activity theory, a concept (such as "anorexia nervosa") has been given many meanings from different groups and theoretical influ-

ences. An individual person acquires many different meanings simultaneous from different contexts, e.g. from psychiatry, from psychoanalysis, from the mass media, and from personal relationships. Such meanings can be more or less in harmony or in conflict. To learn about an object is not in principle to make an arbitrary connection between properties and concepts, but to understand the mutual historical developments of the objects and human practices. Concepts thus have "historical depth" (c.f., Mammen, 1994).

I'll finish this section by saying that the very influential theory about scientific paradigms by Thomas Kuhn (1970) also implies a theory of semantics.

> Kuhn argues against the idea that representations of concepts shift historically from diffuse, unarticulated forms to tightly organised, theoretically driven ones; he suggests, not shifts in the representational nature of concepts, but shifts in which theoretical systems embrace the same, or roughly the same, class of phenomena. ... Thus it is only by analysing concepts relative to theories in which they are embedded that we can decide how components are packaged (cited from Keil, 1989, pp. 20-21).

I find this last sentence important: If we are going to analyse the meaning of concepts, and the relationships between concepts, we have to analyse the theories in which they are embedded. According to influential modern epistemology observations are theory-laden. There are no sharp borders between observations, concepts, and theories. They influence each other, and have done this in a historical process. Therefore we need historically oriented epistemologies to clarify all such "social constructions". In the pragmatic epistemologies such analysis is combined with an analysis of explicit and implicit goals and values. Concepts and meanings are mental tools that we produce to accomplish certain goals. They are instruments to attain (political) goals.

5. Documents and Access Points from a Social Constructivist Point of View

The *form* of a document, e.g. the form of a scientific article, is perhaps regarded as something trivial, and usually regarded as something which has an ideal form, which is final in its historical development, can be standardised, and is independent of content and of epistemological issues. "Publication manuals" exist in most academic disciplines (e.g., *Publication manual of the American Psychological Association*, 4th ed., 1994). They describe in great detail the way articles should be designed. Such manuals have a highly technical and normative character, but they are not reflexive concerning their suggestions in the sense that they do not discuss publication form as an epistemological problem.

Emerging research is beginning to change this view of publication form. This new research is using social constructivism and related theories as the epistemological point of departure. The social constructivistic theory of semantics implies that objects are "social constructs" and meanings are *constructed* in social discourse (most often in ways, that are unconscious for the agents involved). Research articles – as well as other documents – are seen as social constructs and as ways of arguing (but never as the only way).

One of the most influential writers on this topic is Bazerman (e.g., 1988). He traces much of the rhetorical technique in scientific articles back to Isaac Newton (1642-1727). Newton not only discovered the basic laws of macrophysics; he also influenced scientific argumentation and publication for about 300 years. However, nothing remains unchanged, and Bazerman analyses changes in the form, length, and structure of the scientific article in the 20. Century. One of these changes is an increase in the number of references, the nature of cited works, and the distribution of the references within the article. Bazerman's work should be of direct interest to both bibliometric studies and to IR – or rather to a broadening of the perspective of these areas. Bazerman also shows how the publication manual in psychology reflects a behaviouristic point of view, which implies that a manual is not a neutral form, but does reflect some epistemological norms, which can be analysed, discussed and questioned. (In 1995-96 was thus a rather intensive debate in *American Psychologist* about the Publication Manual in this field, e.g., Madigan, Johnson, & Linton, 1995; Madigan, Linton, & Johnson, 1996). The work of Bazerman and other researchers in the area of "composition studies" and "genre analysis" is fruitful for IS not only on the concrete level, but also as an inspiration on the methodological level.

The general conclusion provided by this research is that the structure of documents is being analysed both empirically and theoretically. Such knowledge should be of direct relevance for IR and IS. The composition of documents reflects some epistemological norms, which are often unconscious to the research community. However, these norms can be subjected to epistemological research, and the more or less hidden norms can be discussed or criticised.

6. Parts of Documents and Value Added Elements as Access Points

Almost all the parts of documents and their value-added information (see table 1) have been the objects of research in information science and linguistics.

However, this research is extremely fragmented and scattered and lacks the guidance of better theories.

Titles are important access points, and a rather impressive amount of research has been done on them. Yitzhaki (1996) showed that the ratio of "significant" words varies with discipline and time. Between 1940 and 1990 it rose from 62.7% to 70.2% on the average in the scientific journals which was checked. In the social science journals it rose from 62.5% to 68.9%, and in the humanities from 64.1% to 66.1. One interesting hypothesis discussed by the author is that the availability of KWIC indexes and similar retrieval tools utilising titles might tend to increase the authors' awareness of the importance of titles as retrieval tools. However, no attempt is made by Yitzhaki to illuminate semantic problems in titles, such as metaphoric uses of words. Neither is there any attempt to illuminate what the titles are trying to specify what intended or actual role they have in the communication process. This is briefly shown in Myers (1990), who compares scientific and popular journal titles in biology. However, according to Hjørland, (1997) what should be identified by subject access points is "the informative potentials" of the documents. If this is correct then a more qualitative approach to the analysis of titles is needed.

Abstracts are often – like titles – made by the authors themselves. In IS research is also done in computer abstracting. However, both empirical evidence and theoretical analysis indicate that abstracts in such services as Chemical Abstracts are best made by people knowing the needs of the target groups (see Windsor, 1995, 717-718). The literature about abstracts, abstracting, and abstract journals is very large. Two central sources are Lancaster (1991) and Manzer (1977).

Indexer descriptors and classification codes have semantic problems of another kind than all the natural language fields. A classification system and a thesaurus are (more or less) closed semantic systems, whereas natural languages are open systems. The meaning of "anorexia/de" is established by other rules than is the meaning of "anorexia/ti". The meaning of a class in a classification system or of a descriptor in a thesaurus is in principle determined by *formal* relations to other classes/descriptors and by consistent, internal criteria of application. (In practice, however, a system often applies a given descriptor if the same term appears in, e.g., the title of the indexed document. However, in this case the value of a field consisting in just the copying of information from another field must be questioned).

Table 3: Classification of a Subject Term in an Electronic Thesaurus
PsycINFO (Dialog: Knight Ridder Information, file 11)

e(anorexia nervosa)

Ref	Items	Type	RT	Index term
R1	2963		5	*ANOREXIA NERVOSA
R2	239	B	14	EATING DISORDERS
R3	195	B	7	UNDERWEIGHT
R4	3164	R	4	BULIMIA
R5	794	R	15	NUTRITIONAL DEFICIENCIES
R6	3853	R	32	PSYCHOSOMATIC DISORDERS

?e(eating disorders)

Ref	Items	Type	RT	Index term
R1	239		14	*EATING DISORDERS
R2	16989	B	91	MENTAL DISORDERS
R3	1332	F	1	APPETITE DISORDERS
R4	2963	N	5	ANOREXIA NERVOSA
R5	3164	N	4	BULIMIA
R6	547	N	5	HYPERPHAGIA
R7	3683	N	8	OBESITY
R8	155	R	3	APHAGIA
R9	2845	R	9	APPETITE
R10	262	R	4	BINGE EATING
R11	798	R	5	NAUSEA
R12	794	R	15	NUTRITIONAL DEFICIENCIES
R13	111	R	60	PHYSICAL DISORDERS
R14	39495	R	73	SYMPTOMS
R15	195	R	7	UNDERWEIGHT

A closed system faces a dilemma: It can try to establish consistency within itself. However, when the meaning of a term outside the system is changing then the meaning of the term inside the system seems obsolete. Alternatively, it can try to use the descriptors in agreement with the meaning in ordinary (sub-) language, but then the systems loses its consistency, and the whole idea of having a controlled vocabulary is lost. There are advantages and disadvantages by both open systems and closed systems, and they can supplement each other in IR. What a classification system (or controlled vocabulary) can do is to establish consistency within one collection or database and contribute to some kind of standardisation of terminology. The literature about classification and thesauri is enormous, and no references will be given in this paper. To me it is a question whether it is possible to identify any clear progress in the huge number of papers produced on these issues in the last decades. A link to semantic theories is provided by the technology of "semantic nets" developed in artificial intelligence research and also applied to IR systems (e.g. Sølvberg, Nordbø & Aamodt, 1992).

Introductions are central parts of the documents themselves. Swales (1990) is one of the most influential researchers in this field. According to Malmkjær

(1995, pp. 177-178) his pioneering study was based on the introductions to forty-eight articles, sixteen each from pure sciences, applied sciences and social sciences. After some criticism from other researchers, Swales proposed the following model of the composition of introductions in scientific articles:

Composition of Introductions in Scientific Articles

Move One: *Handling Previous Research*
A: Asserting Importance of the Topic
Or
B: Stating Current Knowledge of the Topic

Move Two: *Preparing for Present Research*
by
A: Indicating a gap
Or
B: Question Raising
Or
C: Extending a finding

Move Three: *Introducing Present Research*
by
A: Giving the Purpose
Or
B: Describing Present Research

This model is but one example of research in genre analysis. It should be obvious that studies of this kind are relevant for developing theories of fulltext IR.

Other elements. Even as special an element as the authors' "acknowledgements" in articles has been subjected to research in IS, which resulted in the publication of a whole book on the subject (e.g., Cronin, 1995). Other special elements such as "key word plus" and "research fronts" (both in the Science Citation Index) have been developed, but so far only subjected to little research (see Garfield, 1990 and Dehart & Scott, 1991).

The general conclusion from this section is that documents (and their value added supplements) consist of many parts which are partly a reflection of cultural norms and partly a reflection of an adaptation to given possibilities and to the communicative needs of authors, publishers and intermediaries. The social constructivist point of view tries to illuminate the historical character of these elements and the hidden assumptions, norms and values in them. Such research is valuable for IR and IS because it uncovers the structures with which information professionals have to work. That such an approach is necessary should be quite obvious, but nevertheless it stands in contrast to the mainstream IR today.

7. References and Citations
(With the Idea of Hypertext-like Knowledge Organisation)

References in scientific documents are listed according to existing standards. Garfield & Small (1997, p. 963) suggest that numbered citations are prevalent among natural science journals, while social scientists prefer the author+year system. The number system is seen as most functional for citation indexing, but Borgman (1995) defends the author+year system. References have become extremely important subject access points since Garfield's construction of the citation indexes (The first of these, the Science Citation Index, started in 1963. See Garfield, 1979). The study of citation behaviour, citation indexing and IR based on citation databases has become one of the most exciting research areas in IS. A valuable reference is MacRoberts & MacRoberts (1989), but the bulk of literature is very large.

From our semantic point of view the basic question is what the semantic relations between a cited article and the citing article are (Cf. Harter, Nisonger, & Weng, 1993). However, some researchers would claim that the relations between cited references are not of a semantic but rather of a pragmatic nature. In my opinion this is a pseudo-question caused by a wrong view of semantics. If we discard theories like "the picture theory", and turn to social theories of semantics, then the meaning of terms are produced in "thought and discourse communities", and these communities are connected to the networks of citing papers.

The relative contribution of citation indexes to IR (compared to term searching) depend both on citation practices, on the explicitness of the sub-language of the documents and on the quality of the indexing systems. The general result of empirical investigations is that term and citation searching supplement each other. More specific guidelines for IR are, however, difficult to establish on the basis of the research done so far. From a social constructivist point of view, citation behaviour is governed by cultural norms, which can be discussed and criticised.

Citation behaviour is extremely important because the goal of IR is to provide the references, which are useful in solving a specific problem. A scientific article is a documentation of the solving of a specific research problem. The problem is formulated in the article, and the documents actually used are cited. Each of the thousands of articles produced daily is in a way a case study in IR. Every article not only poses a definite IR problem, but the list of references provided by the author is the key to how that particular person has solved the problem. Thus it is possible to check theories of IR against this key! Most research on

"relevance" and on IR seems to have overlooked this fact. From what we do know, it seems extremely unlikely that an algorithm should be able to select references from electronic databases and end up with just the set of references represented in a given article. From this point of view, theories of IR seem very naïve and unrealistic. A more detailed study of citation behaviour can illuminate the real problems of IR: That selected documents are not simply a set of documents sharing a fixed set of attributes which are not represented in the non selected items.

Today we do know something about scientists' citation behaviour. Smith (1981, p. 84) mentions fifteen reasons for authors to quote other documents:

1. Paying homage to pioneers
2. Giving credit for related work (homage to peers)
3. Identifying methodology, equipment, and so on
4. Providing background reading
5. Correcting one's own work
6. Correcting the work of others
7. Criticising previous work
8. Substantiating claims
9. Alerting to forthcoming work
10. Providing leads to poorly disseminated, poorly indexed, or uncited work
11. Authenticating data and classes of facts – physical constants, and so on
12. Identifying original publications in which an idea or concept was discussed
13. Identifying original publications or other work describing an eponymic concept or term
14. Disclaiming work or ideas of others (negative claims)
15. Disputing priority claims of others (negative homage)

This list of citer motivations gives an impression of the goals that real IR must meet. It is evident that this is not just a mechanical question, but to a large extent a question implying norms and values. The political character of selecting references becomes even less mechanical, if you consider some of the problems which research has discovered in people's citation behaviour. Seglen (1996, p. 29) thus lists a range of problems concerning selection of references:

1. References are selected because of their usefulness for the author, which is something different from their quality
2. Only a small fraction of all used material is cited
3. General knowledge is not cited
4. Knowledge is often cited from secondary sources
5. Documents supporting an author's arguments are cited more often than other documents
6. Flattering (citing editors, potential referees, and other authorities)
7. Showing off (citing hot new "in" articles)
8. Reference copying (references provided by other authors)
9. Conventions. In biochemistry, for example, methods are cited but not reagents
10. Self citations
11. Citing colleagues (often reflecting informal transfer of information)

This research on citer motivations raises the problem that IR should not only predict what references users would ideally select, but should also be involved in questions about what to regard as ethical citation behaviour, and what to regard as good science! Research in IR cannot escape questions related to the philosophy and methodology of science.

However, this research also says something more technical about the usefulness of references versus descriptors in information seeking: To the degree that the conventions can be described they are of immediate relevance. With the knowledge given above (#9), we are able to state that citation indexing should perform well on a search for biochemical methods, but rather badly on a search for a reagent. Such conventions must, however, be uncovered piece by piece.

"Atlas of Science" is a concept that can be traced to Wilhelm Ostwald, 1919 (cf., Bonitz, 1983). It is based on citation methods, -connections, and -frequencies. They display the connection between research areas such as they can be mapped by cocitation analysis. Garfield (1981) first developed this idea into a concrete (and commercial) product. As tools for IR they share the same kind of semantic problems as bibliographic references.

8. Assessments of Documents and Target Group Analysis

Quality assessments. Documents, which are indexed, are rarely explicitly evaluated in databases. The main evaluation is the selection of journals to be indexed. It is normally assumed that there exists a hierarchy of journals in the single disciplines. However, IR research and system development has not hitherto made any suggestions that such evaluations should be displayed for the user and thus make it possible for him/her to limit a search to sets of highly evaluated journals (or other sources). In a way the value-added services provided by journals are not utilised when searching information in electronic databases.

Sociological Abstracts (SA), does, however, indicate whether book reviews are favourable, neutral, or unfavourable. On January 1998, the following evaluations were registered in SA:

E4 16027 EV=FAVORABLE
E5 3596 EV=NEUTRAL
E6 2909 EV=UNFAVORABLE
E7 974 EV=VERY FAVORABLE
E8 233 EV=VERY UNFAVORABLE

Target group analysis. In professional databases, such as MEDLINE, documents are not classified according to potential user groups. An example of a database doing this is the book review database *"Choice"* published by the Association of College and Research Libraries, USA. The fact that this kind of information is the exception rather than the norm raises interesting questions concerning what "user related" and "cognitive viewpoints" in IR are actually aiming at.

User groups as classified in "Choice"
Users in traditional academic curricula:
Lower division undergraduates
Junior/senior undergraduates
Graduate students
Researchers
Faculty
Users in professional/technical curricula:
Community college students
Preprofessional students
Professionals
Practitioners

9. Interpreting the Result of Database Repackaging

When a user or an information specialist searches databases, he/she has access to a lot of different documents, document representations and subject access points. Each of the possible access points is formed by some human agent (or by a machine programmed by a human agent). Every element has its own history, and it has been formed by some implicit or explicit goals and theories. At the deepest level these theories are the epistemologies uncovered by philosophical analysis.

A given database can be a merging of what were once different separate databases. In the original databases the access points were perhaps not explicit about some aspects of the subject matter because this was implicit in the delimitation of the database. For example, if you search for "lead" in PsycINFO, there is no need to indicate that you are searching the effects of lead on behaviour: it is implicit in PsycINFO that all records are on animal or human psychology/behaviour. However, if the records in PsycINFO are merged with the records in Chemical Abstracts, you will have to change your search strategy and specify that you are searching studies on how lead influences behaviour and performance. This new strategy would probably be less than optimal regarding the part of the records originally indexed in PsycINFO (because implicit information is lost by the merging).

At another level PsycINFO can be seen as a merging of records which were once presented in individual journals, some of which may be American, some European, some behaviouristic, other psychoanalytic, etc. Originally, to the readers of those journals their selection policy and their way of writing titles and composing articles reflected some implicit meanings in those journals. By making a controlled vocabulary, a classification scheme, a certain structure in the records and so on the people behind PsycINFO made certain decisions which were coloured by their view of knowledge. For example, Roberts (1985) showed that most thesauri in social sciences were inspired by natural sciences, and were probably designed according to the principles, which were more suitable for documents belonging to natural sciences than for social sciences for which they were intended. Such (more or less implicit) theories of knowledge can be in harmony or conflict with the (more or less implicit) views of knowledge represented by the indexed journals. Both views can again be more or less in harmony or conflict with the implicit or explicit view of knowledge in the query, which again can be more or less in harmony or in conflict with the user's real information need.

According to modern semiotic theories also the single document should be understood as a merging of several texts. This is called "intertextuality".

10. The Changing Role of the Information Scientists

The job of information science is to contribute to the process of identifying those documents that can be of most value to the user's tasks. It is not possible to formulate a query without any knowledge of what has been produced, in what disciplines/contexts it has been produced, what all the available subject access points are and what are the strengths and limitations of each kind of access points. Therefore *interaction* is such an important concept in IR (cf. Ingwersen, 1992): In modern IR the user interacts with the system and can reformulate the question on the basis of feedback from the system.

In his or her interaction with information systems the user has access to different layers of information provided by different agents: Fulltext or partial texts, abstractor and indexer information, journal or publisher name and database delimitation. These instances can use the same or different words (e.g., "anorexia" or "eating disorders"), and such words can carry meanings in conflict with other meanings.

The most useful information for the user is to know:

1) That a given search term has different meanings
2) A mapping of these different meanings

This must be done through a kind of analysis related to that of "social constructivism", and by digging through layer after layer (by Michel Foucault termed "the archaeology of knowledge"). It is important to realise that the epistemological views of the different layers are often not synchronous in their development. An article in a journal can be written from one epistemology in a journal following principles inspired by a second epistemology. The same article can be indexed in a database influenced by a third epistemology, and used by a user interested in e.g., anorexia illuminated from a fourth epistemological point of view.

Most of the information is not provided by information specialists but by other agents. The job of the information specialists is to make retrieval more efficient. If the system is sufficient without information specialists, there should be no information specialists and no "value added information" provided by information specialists. The most important job of the information specialists is reuse existing information in IR before producing redundant information. The information specialists have to understand the possibilities and the limitations of the information systems from the potential user's point of view.

In the (hypothetical?) cases where there is a high degree of harmony between the meanings of the words in all the different access points, the information scientist only has to make the system user friendly, e.g., by providing some standardisation. It may not be necessary to index the documents because the texts themselves are explicit and sufficient information about their subject matter. Indexing, abstracting, etc. can thus be done by the author of the documents. If IR theories are based on semantic theories related to that of the picture theory, there is not much need to make implicit views explicit. There is no such thing as implicit knowledge (See also Nystrand & Wiemelt, 1991). There is no such thing as different interests influencing concepts and knowledge. IR retrieval looks like a value free mechanical process, and it is hard to see the needs for professional intermediaries when all the necessary algorithms have been developed.

To the degree that concepts can be interpreted differently the illumination of such differences would be of high relevance to the users. Table 3 is a typical example of what is done today in an online thesaurus. It does not map the different meanings of "anorexia nervosa", and in my view this is a major limitation. What would be interesting would be to have a system that could inform the users of the basic theoretical views on anorexia nervosa: Psychiatric/biological theories, psychoanalytic theories, humanistic theories, social and cultural theories and so on. A system that referred to the most influential diagnostic systems such as DSMIV (published by American Psychiatric Association), as well as to criticism of this view and to alternative views. There could be a kind of "artificial intelligence" built into the system in such a way that it would help the user identify the respective journals, other publication forms, concepts, disciplines, geographic localisations, research fronts, etc., in which a particular view on anorexia was represented.

Information scientists may well fear that the suggested approach presupposes more subject knowledge than they possess. The only solution I can see is to approach the problem in a top down fashion starting with the general epistemological theories such as classical empiricism, rationalism, historicism, and modern theories like Kuhn's theory of scientific revolutions. In my opinion IS can be as general a science as can the theory of science and similar fields. However, subject knowledge *is* important, and some degree of specialisation in IS is desirable. This is also institutionalised in IS by separate journals and separate interest groups in, e.g., The American Society for Information Science. It is also important to notice that information scientists with qualifications in epistemology may have a better grasp of such semantic problems than most ordinary subject specialists. Such people are often specialising in very narrow problems and do not have this kind of perspective on information structures. In my opinion there exists a clear need for an IS working along these lines.

In short, what the users need are not "neutral" selections of the documents. Relevance is not a one-dimensional scale based on quantitative properties. Users need "maps" of information structures, which can help them to be oriented and to refine their search arguments. Such maps should reflect the basic approaches and should uncover the more or less hidden meanings, interests and goals in documents.

Information scientists have hitherto been most interested in the standardisation of terminology and they have had an implicit interest in semantic theories related to the picture theory, because such theories seemed to allow for mechanical manipulation. The opposite kind of semantic theories: the pragmatic and interpretative kinds of theories have not so far been attractive for mainstream IS/IR. However, there seems to be a possibility that exactly this kind of theories can motivate a need for information specialists in the future.

Notes

1. An extended abstracts of this paper was presented at The Sixth International BOBCATSSS Symposium in Budapest 26th-28th of Januar 1998: Shaping the Knowledge Society.
2. One of the reviewers of this article (not anonymous to me) wrote: "I do not consider semantics as a fundamental focus of the article. It concentrates on structural components of documents in databases in the context of information retrieval and this is of immediate interest to our readers".
   However, I myself *do* consider this article as a work connecting semantics and Information Science. My inspiration to do this came from Harter, Nisonger, & Wenig, (1993), who described the relationships between cited and citing articles as semantic relationships. What they suggested (and what I have outlined in much more details), is, that from the point of view of information retrieval the relationships between structural components should be regarded as semantic relationships.
3. "Syntactical retrieval" (e.g., chemical retrieval) retrieval in multimedia databases etc. are examples of access points not fitting into the present scheme. However, retrieval with feedback such as Salton's "Smart" do employ such access points (but do not have any theoretical basis regarding their relative role).
4. Wittgenstein was not a member of the Wienna Circle, and not the most influencial person on the semantic theory of logical positivism. This was Rudolf Carnap (1942). However this paper only considers the work of Wittgenstein and should not be considered as a treatment of the theory of logical positivism. Ogden & Richards (1923) is a very important book on semantics bridging the pragmaticism of Peirce and the logical positivism.

References:

Bazerman, C. (1988). *Shaping written knowledge. The genre and activity of the experimental article in science.* Madison, Wisconsin: The University of Wisconsin Press.

Blair, D. C. (1990). *Language and representation in information retrieval.* Amsterdam: Elsevier.

Bonitz, M. (1983): Wie lassen sich die Frontgebiete der Forschung bestimmen? : 'ISI Atlas of Science' für Biochemie und Molekularbiologie. *Zentralblatt für Bibliotbekswesen*, 97(7), 295-296.

Borgman, C. L. (1997). Rejoinder [to: Citation format by Garfield & Small, 1997]. *Journal of the American Society for Information Science*, 48(10), 964.

Brier, S. (1996). Cybersemiotics: A new interdisciplinary development applied to the problems of knowledge organisation and document retrieval in information science. *Journal of Documentation*, 52(3), 296-344.

Brooks, T. A. (1993). All the right descriptors: A test of the strategy of unlimited aliasing. *Journal of the American Society for Information Science*, 44(3), 137-147.

Carnap, R. (1942): *Introduction to semantics.* Cambridge, Mass.: Harvard University Press.

Cronin, B. (1995): *The scholar's courtesy: the role of acknowledgement in the primary communication process.* London and Los Angeles: Taylor Graham.

Dahlberg, I. (1978). A referent-oriented, analytical concept theory for INTERCONCEPT. *International Classification*, 5(3), 142-151.

Dahlberg, I. (1995). Conceptual structures and systematization. *International Forum on Information and Documentation*, 20(3), 9-24.

Dehart, F.E., & Scott, L. (1991). ISI research fronts and online subject access. *Journal of the American Society for Information Science*, 42(5), 386-388.

Dewey, J. (1925). *Experience and nature.* Chicago: The Open Court Publishing Company.

Dewey, J. (1939): *Intelligence in the Modern World. John Dewey's Philosophy.* Ed. by Joseph Ratner. New York: The Modern Library.

Dewey, J. (1946): Peirce's theory of linguistic signs, thought, and meaning. *The Journal of Philosophy*, 42, 383-388.

Ellis, D. (1996). *Progress and problems in information retrieval.* London: Library Association Publishing.

Forrester, M. A. (1996). *Psychology of language: A critical introduction.* London: SAGE Publications.

Foskett, A. C. (1977). Assigned indexing I: Semantics. In: *The subject approach to information* (pp. 67-85). London: Clive Bingley.

Frohmann, B. (1990). Rules of Indexing – A critique of mentalism in information-retrieval research. *Journal of Documentation*, 46(2), 81-101

Garfield, E. (1979). *Citation Indexing: Its theory and Application in Science, Technology and Humanities.* New York: Wiley & Sons.

Garfield, E. (1981): Introducing the ISI Atlas of Science: Biochemistry and Molecular Biology, 1978-1980. *Current Contents*, (42), p.5-13.

Garfield, E. (1990). Key-Words-Plus takes you beyond title words. *Current Contents*, 33(AUG), 5-9.

Garfield, E. & I. H. Sher (1993): KeyWords Plus – algorithmic derivative indexing. *Journal of the American Society for Information Science*, 44(5), 298-299.

Garfield, E. & Small, H. (1997). Citation format. *Journal of the American Society for Information Science*, 48(10), 963.

Hardwick, C. (1971): *Language learning in Wittgenstein's later philosophy.* The Hague/Paris: Mouton.

Harnad, S. (1990). Scholarly skywriting and the prepublication continuum of scientific inquiry. *Psychological Science*, 1, 342-343.

Harnad, S. (1991). Post-Gutenberg Galaxy: The Fourth revolution in the means of production of knowledge. *Public-Access Computer Systems Review*, 2(1), 39-53.

Harter, S. P., Nisonger, T. E., & Weng, A. (1993). Semantic relationships between cited and citing articles in library and information science journals. *Journal of the American Society for Information Science*, 44, 543-552.

Hjørland, B. (1997). *Information seeking and subject representation. An activity-theoretical approach to information science.* Westport, Connecticut & London, England: Greenwood Press.

Ingwersen, P. (1992). *Information retrieval interaction.* London: Taylor Graham.

Karamüftüoglu, M. (1996). Semiotics of documentary information retrieval systems. In P. Ingwersen & N. O. Pors (Eds.), *Proceedings CoLIS 2: Second international conference on conceptions of library and information science: Integration in perspective. October 13-16, 1996* (pp. 85-97). Copenhagen: The Royal School of Librarianship.

Karamüftüoglu, M. (1997). Designing language games in Okapi. *Journal of documentation*, 53(1), 69-73

Keil, Frank C. (1989). *Concepts, Kinds, and Cognitive Development*. London: The MIT Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2rd ed.). Chicago: University of Chicago Press.

Lakoff, G. (1987): *Women, fire and dangerous things. What categories reveal about the mind.* Chicago: University of Chicago Press.

Lancaster, F. W. (1991). *Indexing and Abstracting in Theory and Practice.* London: Library Association.

Lindholm-Romantschuk, Y. (1994). *The flow of ideas within and among academic disciplines: scholarly book reviewing in the social sciences and humanities.* PH.D.-dissertation from University of California, Berkeley. (Available from Dissertation Abstracts International).

MacRoberts, M. H. & MacRoberts, B. R. (1989). Problems of citation analysis: a critical review. *Journal of the American Society for Information Science*, 40(5), 342-349.

Madigan, R., Johnson, S., & Linton, P. (1995). The language of psychology: APA style as epistemology. *American Psychologist*, 50, 428-436.

Madigan, R., Linton, P., & Johnson, S. (1996). APA style: Quo vadis? *American Psychologist*, 51(6), 653-655.

Malmkjær, K. (1995): Genre analysis. In K. Malmkjær (Ed.), *The linguistics encyclopedia* (pp. 170-181). London: Routledge.

Mammen, J. (1994). En realistisk begrebsteori: Om forholdet imellem virksomhedsteorien og den økologiske kognitive psykologi. In J. Mammen, & M. Hedegaard (Eds.), *Virksomhedsteori i udvikling* (pp. 43-58). Århus: Århus Universitet, Psykologisk Institut.

Manzer, B. M. (1977). *The Abstract Journal, 1790-1920. Origin, Development and Diffusion.* Metuchen, N.J.: The Scarecrow Press.

Myers, G. (1990). *Writing Biology. Texts in the Social Construction of Scientific Knowledge. Madison,* Wisconsin: The University of Wisconsin Press.

Nystrand, M., & Wiemelt, J. (1991). When is a text explicit: Formalist and dialogical conceptions. *Text*, 11, 25-41.

Ogden, C. K. & Richards, I. A. (1923): *The meaning of meaning; a study of the influence of language on thought and of the science of symbolism*. New York, Harcourt, Brace & Company, Inc.

Ostwald, W.: Die chemische Literatur und die Organisation der Wissenschaft (in: *Handbuch der allgemeinen Chemie*. Bd. 1. Hrsg von W.Ostwald & C.Drucker. Leipzig, 1919, s.92-).

Peirce, C. S. (1905). What pragmatism is. *The Monist*, 15, 161-181.

Publication manual of the American Psychological Association. (4th ed.).(1994). Washington, DC: APA.

Roberts, N. (1985). Concepts, structures and retrieval in the social sciences up to c. 1970. *Social Science Information Studies*, 5, 55-67.

Seglen, P. O. (1996). Bruk av siteringer og tidsskriftimpaktfaktor til forskningsevaluering. *Biblioteksarbejde*, 17(48), 27-34.

Smith, L. (1981). Citation analysis. *Library Trends*, 30, 83-106.

Stamper, R. (1987): Semantics. In: R. J. Boland Jr. & R. A. Hirschheim (Eds.). *Critical Issues in Information Systems Research* (Chapter 2, pp. 43-78). Chichester: John Wiley & Sons.

Swales, J. (1990). *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

Sølvberg, I., Nordbø, I., & Aamodt, A. (1992). Knowledge-based information retrieval. *Future Generations Computer Systems* 7, (1991/1992), 379-390.

Tuominen, K. (1997). User-centered discourse: An analysis of the subject positions of the user and the librarian. *Library Quarterly*, 67(4), 350-371

van Rijsbergen, C. J. (1986). A new theoretical framework for information retrieval. In American Society for Computing Machinery, *Proceedings of the 1986 ACM Conference on research and development in information retrieval* (pp. 194-200). New York: ACM Press.

Vickery, B. C. & Vickery, A. (1987). Semantics and Retrieval. IN: *Information Science in Theory and Practice* (Chapter 6, pp. 133-179). London: Bowker-Saur. (Reprinted 1989).

Warner, J. (1990): Semiotics, information science, documents and computers. *Journal of Documentation*, 46(1), 16-32.

Windsor, D. A. (1995). Abstract concerns. *Journal of the American Society for Information Science*, 46(9), 717-718.

Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. New York: Addison-Wesle.

Wittgenstein, L. (1958). *Philosophical Investigations*. 3.ed. New York: McMillan.

Woodfield, A. (1991). Conceptions. *Mind*, 547-572.

Yitzhaki, M. (1996). Informativity of journal article titles: The ratio of "significant" words. In P. Ingwersen & N. O. Pors (Eds.), *Proceedings CoLIS2: Second international conference on conceptions of library and information science, integration in perspective, October 13-16, 1996* (pp. 447-458). Copenhagen: The Royal School of Librarianship.

Birger Hjørland, Royal School of Library and Information Science, Copenhagen, 6 Birketinget, DK-2300 Copenhagen S
email: bh@db.dk