Knowledge Based Information Retrieval: A Semiotic Approach

Volume 1: Theory, Systems Design and Evaluation

H. Murat Karamüftüoğlu

Thesis submitted for the degree of Doctor of Philosophy

Department of Information Science
City University
London

June 1998

Table of Contents

Volume 1: Theory, Systems Design and Evaluation

Volume 2: Appendices

List of Figures

Acknowledgements

I wish to express my gratitude to all (past and present) academic and administrative staff of the department, in particular to Stephen Robertson, Mike Gatford, Stephen Walker, Thien Do, and Susan Jones for their support, patience and advice. This work is dedicated to my father Ahmet Karamüftüoğlu..

Ahfadımın en son doğacak ferdine benden
Bir tuhfe-i iman götür ey son nefesim sen!

Süleyman Nazif

Declaration of Copyright

Abstract

The overall objective of this study is to analyze the document retrieval process and the main information retrieval (IR) concepts from the point of view of semiotics and design retrieval mechanisms based on the findings of the semiotic analysis of the retrieval situation. Semiotics is a discipline which studies 'sign systems' and how signs are exchanged in communication. The semiotic view of IR interaction presented in this dissertation views document retrieval as a kind of human communication process taking place in a social and cultural realm.

The most important result of the semiotic model developed is the explication of the distinction between the knowledge production and transfer functions of document retrieval. The consequence of this finding is the conceptualization of the retrieval process as a dynamic and complex interplay between knowledge production and transfer tasks. It is hypothesised that, in the case of knowledge production, users of retrieval systems are interested in exploring new areas of the document collection which are not *a priori* known.

Two knowledge based systems are developed based on the Okapi probabilistic retrieval system. The purpose of the retrieval systems designed is posited, in general terms, as to suggest the users new search areas of potential interest. This is achieved by treating the Inspec thesaurus as a semantic network, and applying a heuristic spreading activation technique to generate clusters of terms linked in the Inspec thesaurus. Each cluster or batch of terms is conceived as representing a part of the general search area defined by the initial user search terms. The main design objective here is to enable the user to identify new search areas from the term information contained in the batches.

Two evaluation experiments were carried out with real users who had real information needs to test whether the batches were actually effective in defining search areas related to the original user queries and whether they were useful in pointing new areas which were potentially relevant to the users. A number of hypotheses related to the retrieval effectiveness of the knowledge based systems designed were also tested in the experiments. The main findings of the experiments indicate that:

- the batches were useful in representing search domains relevant to the users' queries

- in many cases the batches represented new ideas or new search domains to the users

- the knowledge based systems had similar retrieval effectiveness in terms of precision as the Okapi system

Chapter 1
Introduction

The purpose of Information Science, in general, and Information Retrieval, in particular, is to facilitate the communication of information between human beings (cf. Belkin & Robertson, 1976). This statement has been expressed by various authors in slightly varying forms at different times (see e.g. Vakkari & Cronin, 1992, Ingwersen & Pors, 1996). This statement is also at the inception of the present dissertation. It can be said that, the present dissertation has started by accepting the above statement.

When the above statement is accepted, the next logical step is to analyze the information retrieval process from the perspective of an appropriate discipline that studies human communication phenomena. There are several disciplines that study human communication from different angles, such as, media and communication studies, cultural studies, cognitive science, semiotics. From a more general perspective, psychology, sociology, aesthetics, linguistics, philosophy are all involved with some aspects of human communication.

In this study, semiotics has been taken up as a general methodological framework in the analysis of information retrieval (IR) as a kind of human communication process with a view to build retrieval methods/systems based on the semiotic analysis. There are several reasons for this choice. One of the reasons is that, semiotics has been widely used in other areas related to human communication with some success (see chapter 3). More importantly, semiotic analysis includes social aspects of the communication process and in this capacity has the ability to establish contacts with disciplines that study social issues, such as, anthropology, sociology (including sociology of science), politics, economics, and philosophy (cf. chapter 3). A third reason is related to the previous one. Although, document retrieval[1] has been studied extensively from individualistic points of view (e.g. from cognitive and psychological perspectives), it has seen relatively little so far in the way of rigorous and thorough analysis from more socially oriented perspectives. It is assumed here that, analysis of information retrieval process in a broader social context is both an appropriate and indeed timely endeavour in the light of advances in networked information systems, which allegedly dissolve the traditional boundaries between disciplines. It has also been noted that, traditional models of information seeking behaviour based on the idea of an individual user searching a library catalogue has become unfruitful in the face of emerging networked services and fracturing of the information-scape (see e.g. Cronin & Hert, 1996).

Although high quality research has been emerging lately which applies semiotics to computer systems (see e.g. Andersen, 1990; see also 3.1), so far this research has not been extended to information retrieval. There has been recently some important work which analyze document retrieval from the perspective of disciplines closely allied to semiotics, such as language games and speech acts theories (see 2.3), however, there has not been so far a thorough and detailed study which employs rich and varied conceptual tools offered by semiotics. This can be considered both as a challenge and a disadvantage.

The challenge is to carry out a detailed investigation of the retrieval situation from the

[1] Document retrieval and information retrieval are used interchangeably through out this dissertation.

perspective of semiotic theories and apply the results of this investigation to design new retrieval methods. The disadvantage involved in this challenge is that, there is no solid groundwork from which one can proceed. As a result of lack of firm foundation from which a semiotic analysis of the retrieval situation can be conducted, I had to start in the present study from the very fundamentals of the retrieval process and apply the basic semiotic concepts one by one in order to build an understanding of information retrieval that conforms with semiotic view of the human communication processes.

One of the consequences of building the semiotic model of document retrieval from scratch is that, a substantial part of the present dissertation had to be allocated to detailed analyses of some of the basic concepts used in information retrieval theory (such as, information transfer, relevance, sign, etc.). As a result, some of the analyses of the document retrieval process have to be performed at a relatively high level. Another challenge follows from this last point. Since the semiotic analyses have to be carried out at a relatively high level (as there is no previous work on which one can build a more detailed model), it was a challenge to relate the semiotic model developed to the tasks of actually building and testing retrieval systems, which are the ultimate aims of the present project.

The approach taken in this project is to perform, first, a detailed theoretical analysis of document retrieval as a human communication process from scratch, and based on the findings of this analysis attempt to build new retrieval methods which conform with the main findings of the theoretical analysis. It should be noted that, some tools are taken as a *given*. The Okapi retrieval system was a given tool which served as an application platform. Similarly, the Inspec thesaurus encoded as a relational database was available and used as a knowledge base. No attempt has been made, for instance, to analyze or criticise the effectiveness of Inspec as a thesaurus. These are all taken as givens. The challenge was, given such tools how to design retrieval systems which reflect the theoretical model of IR developed. The theoretical model of IR based on semiotics, therefore, has served as a guide which illuminated the systems design and evaluation practices by clarifying the principal concepts involved in IR theory.

The most general objective of the present project can therefore be said to answer the following question: given some retrieval tools (namely the Okapi retrieval system and the Inspec thesaurus), how to go about building retrieval systems based on the semiotic analysis of the IR situation?

One of the main findings of the semiotic analysis carried out as part of this project is that, there are two distinct functions of document retrieval systems, namely, *information* or *knowledge transfer* and *knowledge production*. It is subsequently hypothesised that, some of the users of the Okapi system might be involved in a knowledge production activity and as result of this, would like to explore areas of the document collection which are not known to them. The assumption here is that, the users have a general area of interest as implied by their search terms, however further specification of potentially useful search areas is not available (since this cannot be known *a priori* in a knowledge production setting).

As a result of the above analysis, the purpose of the retrieval systems designed is posited, in general terms, as to suggest the users search areas of potential interest which are not implied directly by their search terms. This is done by extracting terms from the Inspec thesaurus related to users' search terms and presenting the extracted thesaurus terms in small conceptually related clusters or batches. Semiotic analysis of the retrieval situation was instrumental in deciding to use batches (of terms linked in the thesaurus), rather than single terms in the knowledge based systems designed (cf. 6.4.2). An important design decision was to include redundant terms (not closely associated with user's search terms) in the batches in order to direct the search to areas not immediately obvious from the original user input terms. Inspec thesaurus is treated as a knowledge base with a *semantic network* type of knowledge representation scheme and a heuristic based *spreading activation* technique is devised to generate batches of linked thesaurus terms. Two evaluation tests are carried out with real users who had real information needs to test whether the generated batches are useful in suggesting new ideas to the users, as well as number of other hypotheses and questions (see outline of the individual chapters comprising the present report below).

The present dissertation is organized along two main parts. The first part comprise of reviews of relevant literature (chapters 2, and 3) and semiotic analysis of the retrieval situation (chapters 4, and 5). Chapter 6 serves as a link between the first part and the second, and includes summary of the semiotic model developed in the first part and discussion of how the theoretic model is used in designing and evaluating systems in the second part. This chapter also presents formulation of the design objectives and hypotheses tested in the evaluation experiments. The second part of the dissertation comprise of detailed description of the systems designed based on the semiotic model developed (chapter 7) and results of the evaluation experiments performed (chapter 8). The concluding section of the report (chapter 9) discusses the general results of the project and make recommendations for future research.

In the remaining part of this introduction contents of each of the chapters are outlined in little more detail.

Chapter 2 presents an overview of the retrieval methods and systems which have a direct bearing on the systems developed and evaluated in this project. These include, probabilistic models of IR, relevance feedback, term clustering and query expansion methods and knowledge based approaches to IR, especially, those based on semantic network and spreading activation techniques.

Chapter 3 introduces the basic concepts used in semiotics. There are several schools of semiotics with a rich variety of concepts and tools. The review in this chapter is mainly concentrates on the post-Saussurean and Peircian varieties of semiotics, including contributions of Eco, Barthes, Hjelmslev and others.

In chapter 4, a syntactic analysis of information retrieval systems is performed. This analysis includes identification of the types of signs, expression and content planes in IR, and types and levels of coding in retrieval systems. This type of analysis is sometimes called as 'semiotics of signification' or 'theory of codes'.

In chapter 5, a 'theory of sign production in IR' is presented. This chapter presents a discussion of the modes of sign production in information retrieval, and propose a communication model for IR. This type of analysis is sometimes called as 'semiotics of communication', or 'theory of sign production'.

Chapter 6 summarizes the semiotic model of information retrieval developed in the preceding chapters and applies the main principles derived from this model to systems design and evaluation tasks. In this chapter the main objectives of the systems designed are formulated (6.3 and 6.4) and hypotheses and questions to be tested in the evaluation experiments are determined (6.4.3). The main hypotheses explicated in section 6.4.3 are related to the effectiveness of the batches of linked thesaurus terms in representing user queries, and suggesting new search areas to the users, as well as the effectiveness of the batches in retrieving relevant documents.

Chapter 7 describes the details of the knowledge based systems designed and tested in this project. The Inspec thesaurus which served as knowledge base and its implementation in Oracle database are described. Rules and heuristics used in the systems designed are also presented and discussed.

Chapter 8, describes the experimental procedure followed in the evaluation of the knowledge based systems. This chapter includes a summary of experimental methods in IR and presents a detailed analysis of the results of the experiments performed.

The final chapter, chapter 9, summarizes the conçlusions drawn from this study and points to future research directions and questions.

Chapter 2
An Overview of Probabilistic, Language and Semiotics Oriented Approaches to Information Retrieval

The purpose of this chapter is to review the literature on selected methods and approaches to information retrieval which have direct relevance for the development of the arguments in the rest of the present dissertation. Therefore, the discussion in this chapter does not aim to be exhaustive. Only systems and approaches to information retrieval that have explicit bearing on the project presented in this dissertation are reviewed.

The chapter is divided conveniently into three parts. In section 2.1, information retrieval systems, models and methods that are relevant for the discussion of the approach and systems developed in this project are reviewed. Section 2.2 discusses the use of thesaurus in information retrieval. In section 2.3, literature on the relationship between semiotics and information retrieval are discussed and summarized.

2.1 An Overview of Probabilistic and Language Based Approaches to Information Retrieval

In this section, probabilistic and language based approaches to information retrieval are reviewed. The selected approaches are those that have a direct bearing on the present project. In section 2.1.1 probabilistic and related approach to IR are briefly discussed. This is followed by a review of systems that use best match searching and various methods of query expansion (2.1.2). In section 2.1.3 natural language processing in IR is briefly discussed.

2.1.1 Probabilistic information retrieval

The systems designed and evaluated as part of the present project make use of to a large extent the probabilistic model of the Okapi system (see 7.2 and 7.3). Documents marked as relevant by the users constitute an important source of information. This resource is used by probabilistic systems to expand the user's original search terms which are then used to search for more relevant documents. The above and related points are discussed below.

Probabilistic approaches

The task of information retrieval systems, as generally accepted, is to present the user with the texts which are most likely to satisfy the user's information need based upon the request put to the system (see e.g. Robertson & Belkin, 1978a).

However it can be hypothesised that, there is always a discrepancy between stated request for information and the latent information need underlying the request for information or the query statement. It is because of this discrepancy that probabilistic concerns enter into information retrieval. This view is expressed by Robertson & Belkin (1978a, p.96) as follows: "... we take it as axiomatic that there will be discrepancies between requests and needs, and thus that some probabilistic ideas must enter into retrieval".

The above probabilistic view states that, "if the user could state his/her need completely and exactly, and if the indexing of the texts were also complete and exact, then the probability concept would not arise: perfect retrieval would be possible" (ibid). However, because of the incomplete information available to the system in real retrieval situations, this is not normally possible. The probabilistic approach described states that, in these circumstances "... the system should respond by ranking the texts in order of their *probability of relevance* (according to information available to the system)" (Robertson & Belkin, 1978a, p. 94). It should be noted that the above probabilistic view also assumes a dichotomous relevance variable (ibid), i.e. a text is either relevant to a need or is not (see also 6.1.2).

Ideal response of an IRS according to the above view is not a ranked list but the exact set of documents which will be judged relevant. However, as this is not possible because of the reasons discussed above, an IR system, according to the above probabilistic approach, should instead optimize its performance by ranking the documents in decreasing order of their estimated probability of relevance to the user's need or query. This view is formally expressed by Cooper: "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data" (quoted from Robertson, 1977a, p. 295). This is known as the Probability Ranking Principle (PRP; see 6.1.1 and 6.1.2).

However, there are cases where PRP does not yield to optimal results. Basically, PRP works document-by-document, i.e. documents in the collection are treated independently of each other, and each separate instance of a request is evaluated independently. When the objective is to rank documents in response to a given request (i.e. class of users who put the same query to the system), PRP may not lead to optimal ranking of documents (Robertson, 1977a).

Since, PRP requires binary relevant judgements, in cases where degree of relevance of a text to the user's needs matters, PRP is also not applicable (Robertson & Belkin, 1978a). Therefore, ranking of texts according to probability of relevance (PRP) should not be confused with ranking of texts according to *degree* of relevance. However, in many theoretical and experimental works the two principles have been frequently mixed up which causes confusion with regard to the objectives of the retrieval systems and evaluation exercises (Robertson & Belkin, 1978a).

Degree of relevance has nothing directly to do with the formal statement of the user's need, nor with the retrieval mechanism, but concerns the relationship between text and information need. It is hypothesized that the user will judge some documents to be more relevant in satisfying the need than the others (Robertson & Belkin, 1978a, p. 96).

While it should be desirable to combine both principles to arrive a single *matching function* which makes predictive statements about relevance or degree of match between a document and a request according to both principles, it is shown by Robertson & Belkin (1978a) that, there exists no obvious way of achieving such an objective.

The PRP assumes term and document independence, however, there are other probabilistic models which takes into account the inter-term dependence. Some examples of such work are: van Rijsbergen (1977); Harper & van Rijsbergen (1978); Bookstein & Kraft (1977). It can be noted here that, term and document clustering techniques by definition make use of term/document dependence information (term clustering is discussed in the next section).

The PRP has been put in direct use to device a probabilistic *searching* formula by Robertson & Sparck Jones (1976). This probabilistic model of searching is discussed further below. Some other probabilistic models concern different aspects of the retrieval process, such as, indexing. An earlier example of a probabilistic model of *indexing* is the work of Maron & Kuhns (1960). This model assumes that search terms are *given* (i.e. the user describes the her query irrespective of the system), and the objective of the system is to ensure that the documents are indexed appropriately so that the system would be able to retrieve the relevant documents that will satisfy the user using those particular search terms. This probability of relevance are reflected on the 'relevance number' or weights assigned to indexing terms.

The probabilistic searching model on the other hand, assumes that document representations (index terms) are *given*, and it is the function of the system to ensure the optimal weighting of the search terms (Robertson, 1994). These two models are in fact dual to each other (ibid), and a unified model which incorporates both probabilistic indexing and searching has been put forward by Robertson, Maron & Cooper (1982; 1983).

Other models of probabilistic indexing have also been proposed, most notably by Bookstein & Swanson (1974; 1975), Harter (1975a; 1975b), Salton et al. (1981), and Fuhr & Buckley (1989).

Probabilistic searching

The Probability Ranking Principle discussed briefly above has been used directly to derive a matching function by Robertson & Sparck Jones (1976). The probabilistic model presented in this paper is based on term occurrences in document representations, and works by assigning weights to search terms. Therefore, this is an example of a probabilistic model of searching.

An important aspect of Robertson & Sparck Jones' model of searching is the incorporation of the relevance feedback information into the retrieval process. The relevance weighting theory of Robertson & Sparck Jones is based on the idea that the higher the odds of a term appearing in a document known to be relevant, the higher the odds that the same term will occur in relevant documents which have not yet been seen by the user. The following formula is used to assign weights to terms in the documents marked as relevant by the user (a more detailed description of this formula is given in section 7.2.2):

$w_t=\log\{(r+0.5)/(R-r+0.5)\}/\{(n-r+0.5)/(N-n-R+r+0.5)\}$

where;

wt: is the weight for term t

n: is the number of documents in the database (collection) containing the term t

N: is the number of documents in the database

R: is the number of documents chosen as relevant

r: is the number of relevant documents containing the term t

The terms from the relevant documents are then used to search for more relevant documents (see 7.2.2 and 7.3.1.3 for more details). This process is usually referred to as 'automatic query expansion' or AQE (see 2.1.2).

At the initial iteration when there is no relevance the above formula reduces to:

w_t=log{(N-n+0.5)/(n+0.5)}

Croft & Harper (1979) suggests the following formula should be when there is no relevance information:

w_t= Constant + log{(N-n)/(n)}

which is virtually identical to Robertson/Sparck Jones formula if the constant is set to zero.

Sparck Jones (1972) on the other hand proposes a formula which is shown empirically to be an effective search device:

w_t=log(N/n)

This formula assigns weight to terms according to the number of documents in which they occur in a collection, and known as the *inverse document frequency* (IDF) or *inverse collection frequency* weighting. This formula can be used to calculate weights of terms when there is no relevance information. In fact, as n is normally very much smaller than N, the second part of to the Croft/Harper formula is almost identical to Sparck Jones IDF weight (Robertson & Walker, 1997).

The above Robertson/Sparck Jones probabilistic model of searching is the basis of the Okapi retrieval system which is described in detail in section 7.2 (see also Robertson, 1997). The same model has also been used to develop the CIRT prototype system (Robertson & Thompson, 1990; Robertson et al., 1986) which is used as a front-end to a traditional Boolean retrieval system, namely Data-Star.

There are other systems based on variety of models of probabilistic retrieval. Works of Turtle & Croft (1990), Wade et al. (1989), Wade & Willet (1988) and Croft & Thompson (1987) are some notable examples of such systems. It is also worth nothing here that, there are several other system that use variety of methods to weight search and index terms, most notably the SMART system which uses the vector space model (Salton, 1971; Salton & McGill, 1983). This system and number of others which use best (partial) match searching and relevance feedback mechanism are discussed in section 2.1.2 below.

2.1.2 Best match systems, automatic/interactive query formulation and expansion

Post-co-ordination searching, and more generally Boolean logic, have been important developments in IR history towards the design of effective information retrieval systems. One limitation of Boolean searching however is that, documents whose representations that do not match *exactly* the representation of the user's query in Boolean logic are likely to be missed out.

It is preferable instead to rank the documents according to their degree of match with the query[1]. Probabilistic methods reviewed in 2.1.1 aim to do just that. However, there have been other methods proposed and used in IR which use various term weighting schemes to rank documents according to degree of match. Such systems are generally known as best match or partial match systems.

The simplest weighting scheme is where each term has equal weight and the documents are ranked according the number of terms in common with the search terms (Braekevelt & Wade, 1995). This is also known as 'quorum' or 'coordination' matching. Sparck Jones' IDF weighting scheme described earlier is another example of best match retrieval techniques.

One of the most popular best match scheme is, however, the 'vector space model' (Salton, 1971; Salton & McGill, 1983). In vector space model, documents and queries are represented as weighted numeric vectors. Documents are retrieved by matching the query vectors against the document vectors. The query and document weights are calculated by computing term frequencies in documents and in collection. Terms are treated as independent items. The queries and documents are matched by using a similarity measure known as the *cosine correlation*. The vector space model which lacks the firm theoretical foundation of some of the probabilistic models (e.g. Robertson/Sparck Jones model discussed earlier) has been extensively used, in particular in the Smart system (Salton, 1971; Salton & McGill, 1983). The system incorporates a relevance feedback mechanism and the user's original query is modified by adding term vectors for all retrieved relevant documents and subtracting term vectors for all retrieved non-relevant documents. This results, in general, many new terms added in the query, and the weights of the query terms are adjusted after each iteration. This process is generally known as Automatic Query Expansion (AQE) or query re-formulation.

Term Clustering

Query expansion was subject to great deal of research in the late sixties and early seventies. At that time, query expansion research was mainly concentrated on term clustering methods (Smeaton & van Rijsbergen, 1983; Efthimiadis, 1992).

Term clustering is the operation of grouping together semantically related indexing terms (Lewis, 1992). In cluster analysis groups of objects which have similar values for some set of features are grouped (Lewis, 1992). In term clustering the objects of cluster analysis are index terms which are themselves are features of documents. In most term clustering techniques documents are used as *metafeatures* to group similar terms (ibid). Presence and absence of a term in metafeatures (mostly in documents) are used to group similar terms. Therefore, for a collection of 200 documents each term would be represented by 200 metafeatures, each metafeature indicating presence or absence of the terms in one of the 200 documents (Lewis, 1992, p. 38).

The effects of term clustering on retrieval have been extensively studied, in particular by Sparck Jones (1971; 1973; Sparck Jones & Jackson, 1970). The term clusters are formed based on co-occurrence of single word stems in documents. Her results indicated that small clusters of low frequency terms are most effective, regardless of the clustering method used, however retrieval effectiveness actually improved only on one small collection.

Peat & Willet (1991) have recently pointed out the limitations of term co-occurrence data for

[1] Even Boolean retrieval can be seen as a two-position ranking placed on the entire collection of texts (Robertson & Belkin, 1978a).

automatic query expansion. They demonstrated that clustering terms with documents as metafeatures tends to produce clusters of terms that have comparable frequencies of occurrence, which is not necessarily desirable (Lewis, 1992, p. 38).

Lesk (1969) argued that in small collections term associations only capture local meanings of terms, and do not reflect their general meanings in technical texts. Another conclusion was "... a properly made thesaurus is generally preferable to associative methods" (ibid, p. 322).

Automatic/interactive query formulation and re-formulation (expansion)

The process of modifying the user's original query by adding terms from some source of structured information such as thesaurus, or from documents known to be relevant to the query, is known as query modification, expansion or re-formulation. The original user search terms may or may not be included in the modified query.

The process can be totally automatic, where the system decides which terms are added and/or subtracted from the query, or it can involve the user in term selection process. The former process is usually referred to as automatic query expansion (AQE), the latter interactive query expansion (IQE).

Relevance feedback and term clustering mechanisms discussed earlier are some of the tools used in query expansion process. In the case of relevance feedback terms are extracted from relevant documents for inclusion in the expanded query. The Okapi system based on Robertson/Sparck Jones probabilistic model, and the Smart system based on the vector space model are two examples of systems that use relevance feedback for query re-formulation. Both of these systems use usually free-text terms to modify the user's query. However, other sources of terms are also used, in particular the terms from descriptor (or subject headings) fields of documents. An example of a system using controlled vocabulary for query expansion is CITE (Doszkocs, 1983), which is discussed with other thesaurus based systems in section 2.2.

Term clustering, on the other hand, relies on mainly statistical term co-occurrence data to calculate term to term similarity information. Terms similar to user input terms found in this way are then used to modify the original query.

Another important source of term similarity information is conventional thesaurus. Thesaurus is used by various researchers to automatically or interactively expand users' queries. The project reported in the present dissertation also makes use of information embedded in a thesaurus to re-formulate users' queries. Review of systems using thesaurus and similar structures in the query expansion and retrieval processes is presented in more detail in section 2.2 below.

There are also systems that assist query *formulation*. The objective here is to help the searcher in the initial formulation of the query and/or the search statement submitted to the system. Human end-users or intermediaries of traditional retrieval systems often consult thesauri, classification schemes, dictionaries or lists of subject headings to structure their queries and represent the concepts present in the queries according to indexing rules of the retrieval system. Some of these and other functions involved in the retrieval process are automated to various degrees using rule-based and other techniques.

TomeSearcher (Vickery & Vickery, 1992) assists the users before going online in tasks such as, database selection, query formulation by incorporating terms from a thesaurus, and mapping user input natural language terms into Boolean logic. Shoval (1985) reports another system which assists users in the query formulation process by suggesting terms from a thesaurus. CANSEARCH (Politt, 1987) and MenUSE (Politt, 1988; Smith et al., 1992) assist users in formulating their queries by suggesting terms from the MeSH classification scheme. In the case of MenUSE, the users do not type search terms, instead use a menu driven mechanism to browse and select terms from a list of subject headings. The front-end system designed for the CILKS system (Jones et al, 1995; Jones, 1993; 1992) assists the users in selecting terms from the Inspec thesaurus to formulate their query.

While the above systems involve user interaction in the query formulation process, some others automate the query formulation and related functions. EXPERT (Yip, 1981), and CONIT (Marcus, 1983) use expert system techniques to map the users' queries into Boolean strategies, select appropriate databases to search and carry out the search automatically. The CIRT system (Robertson & Thompson, 1990; Robertson et al., 1986) automatically translates user input search terms into a series of Boolean searches. These are not meaningful Boolean statements but just a means of conducting weighted searching and document ranking on Data-Star which allows only Boolean searches and set retrieval.

A more detailed survey of various systems and front-ends developed to assist users at various stages of the retrieval process can be found in Vickery & Vickery (1993), and Efthimiadis (1990).

2.1.3 Natural language processing

Natural language processing (NLP) and information retrieval share the same research object, namely, text, and features of text, such as, terms, phrases, and larger structural parts (e.g. paragraph). A number of different activities are collectively constitute what is commonly referred to as information retrieval. These activities include, document (text) retrieval, text routing (or selective dissemination of information), text categorization, document and term clustering, and term categorization (Lewis, 1991).

NLP on the other hand, speaking broadly, involves all computer based approaches to handling unrestricted written and spoken texts, and includes such applications as, extracting formatted data, answering questions, and abstracting and summarizing documents (Doszkocs, 1986; Lewis, 1991). Many of the techniques in NLP can be and some of them are applied to various tasks related to document retrieval process, in particular, stemming, automatic or semi-automatic indexing of documents using multi-words or phrases, and query formulation (e.g. mapping of natural language queries to Boolean statements).

It has been suggested by some researches that, IR problem (i.e. presenting to the user the documents that will satisfy her/his information need) could be solved by using NLP techniques to understand what queries and documents really mean (e.g. Lewis et al., 1989). However, it is generally agreed that, it is not expected in the foreseeable future that NLP techniques will develop to the point where deep semantic analyses of documents and queries could be performed.

However more importantly, it is questionable from the point of view of the semiotic perspective developed in the present dissertation whether this is even a desirable thing. It is hoped that, semiotic analyses of chapters 4, 5 and 6 will show that interpretation of documents and production of new knowledge by linking documents in hitherto unknown ways are important aspects of the retrieval process and therefore unequivocal interpretation of documents or queries may not be expected in may cases. In certain situations, it might well be desirable to construct queries ambiguously (i.e. broadly) in order to retrieve documents that are not similar to the documents already known in a subject area, as argued in the later parts of this dissertation. Furthermore, as argued by Hjørland (1997), meaning of a document involves epistemological level of analysis, and understanding a document involves evaluation of epistemological assumptions implicitly or explicitly expressed in the text. This conclusion makes clear that understanding documents and queries involves much more complex issues than it is generally recognized in NLP, and often involves pragmatic as well as epistemologic components. Therefore, even if performed at a deep semantic level, NLP techniques would not necessarily solve problems involved in IR.

An interesting application of NLP from the point of view of the present project, is the analysis of natural language texts to identify various relationships among the linguistic units. This operation when applied to user input natural language texts is sometimes referred to as *query processing*. The objective of this operation is usually to translate user input natural language query into search terms represented in an intermediate query language (Vickery & Vickery, 1992; 1993). Since, in this project mapping of user input free-text search terms into thesaurus terms is performed (see 7.3.1) for query re-formulation purpose it is worthwhile to spend some time on the query processing research.

The aim of query processing according to Vickery & Vickery (1992; 1993) is to achieve unambiguous representation of the internal meaning of the query. Quoting from Vickery & Vickery (1993, p. 121): "This analysis must:

a. eliminate material in the input that does not contribute to the meaningful content (e.g. words that are normally put into a stoplist; unnecessary morphological variants as plural words);

b. disambiguate words with multiple meanings;

c. recognise semantic relations among the remaining words (e.g. form appropriate compounds from adjacent words)".

To carry out the above outlined task analyses of the natural language input at morphological, lexical, syntactic, and semantic levels need to be performed.

The morphological level involves processing of the text at individual word forms level and identification of prefixes, infixes, suffixes and compound words. The lexical level deals with operations on full words, such as identification of stopwords, and misspellings, handling of acronyms and abbreviations, and assignment of parts of speech categories to lexical items. The syntactic analysis of natural language texts deals with recognition of structural units, such as noun phrases. The semantic level of analysis involves representing the meaning of the natural language text by adding contextual information to the syntactic analysis (Doszkocs, 1986).

In order to perform NLP at the above mentioned levels, it is necessary to have both processing rules and a lexicon, which may include a Machine Readable Dictionary (MRD) and/or a thesaurus (Vickery & Vickery, 1993).

Mostly, morphological, lexical and syntactic analyses are attempted in IR research. An example of the analysis of both queries and documents at the syntactic level can be found in the work of Salton and his colleagues (Salton et al., 1990). These researchers attempt to identify multiword phrases, and syntactic variants that refer to the same underlying concept. Heads and modifiers of clauses are also identified and distinguished in the matching operation. Several other systems employ variety of techniques to process user queries at various levels. These systems mostly employ a thesaurus or some other structured file of words, e.g. MRD. Some of these systems are described below in greater detail in the discussion of the role of thesaurus in query formulation and expansion (sections 2.2.1 and 2.2.2).

It is worthwhile to note here that the Okapi system which provided the platform for testing the ideas developed in this dissertation does not require from the searchers to conform with the normal syntactic/semantic rules of natural language discourse to describe their queries (cf. 7.3.1.1). As Okapi is a best match system (7.2), users often use a few keywords to describe the sort of documents that they want to retrieve. In the process of entering the search terms, duplicate terms are often eliminated and slightly varying forms of the same word are sometimes repeated. Thus, the resulting search statements are in fact lists of terms without much structure and usually not suitable for formal analysis using NLP techniques. If it was possible or practical to use NLP techniques to determine noun phrases that may present in users' queries, this information could then be used to map the query terms onto thesaurus terms for the query expansion (re-formulation) purpose. A more suitable method in the case of Okapi searches is to relate the user's query as a *whole* to the thesaurus terms (see 2.2.1 below, and 7.3.1).

2.2 Thesaurus as a Knowledge Source

Thesaurus has been traditionally an important tool in information retrieval. There are numerous volumes on the use of thesaurus in IR (an example is the volume by Lancaster, 1986). The main use of thesaurus in the search process is seen traditionally as bringing variants of the same underlying concept to the attention of the searcher.

A more careful analysis would reveal that, thesaurus and similar vocabulary control devices constitute a (closed) language system on their own right, which are used to represent the contents of documents. Although, terms constituting a thesaurus resemble natural language words and phrases, their meaning in a thesaurus are often quite different from their meaning in natural language discourse (cf. 3.5.9, 3.6). In other words, a thesaurus term attracts and repels usually different sorts of terms than a similar term found in natural language discourse, therefore, it can be said that a term assumes different meaning(s) inside and outside of thesaurus.

Searchers often consult thesauri to translate the concepts present in queries into their correct representation in the indexing language of the retrieval system. In terms of the formulation of search tactics, a thesaurus is used to broaden or narrow the query by moving up or down in appropriate hierarchies that arrange terms by the genus-species relationship. The associative relationship, on the other hand, is usually used for finding concepts semantically related to those present in the query.

Researchers, such as Brooks et al. (1986), have studied the role of the intermediary in the search process in detail. Harter & Peters (1985) suggest useful general heuristics which include actions taken with thesauri in search term selection. Bates (1979) and Fidel (1985; 1986) present useful accounts of search heuristics for Boolean systems. Some of these functions can be automated to varying levels of success and numerous systems and front-ends have been devised to achieve this end. Some of these systems have already been mentioned in the preceding sections. In the following section, some of these systems will be described in more detail. Section 2.2.1 summarizes the methods and systems used in automatic/semi-automatic query formulation and expansion. Section 2.2.2 looks at a specific application of thesaurus in the search process:

thesaurus as a semantic network. Finally, section 2.2.3 summarizes the research done in the area of automatic construction of thesauri and similar knowledge bases.

2.2.1 Automatic/Interactive query formulation/expansion using thesauri

One of the first systems which assists user in the query formulation process was CITE (Current Information Transfer in English). CITE (Doszkocs, 1983; Doszkocs & Rapp, 1979) is a front-end for the US National Library of Medicine's online book catalogue CATLINE. The system allows users to express their queries freely as natural language sentences or term lists. User input query is scanned and broken into words. After words in the stoplist are removed, remaining words are stemmed and matched against dictionary terms and MESH headings. The identified terms, including the MESH headings are then weighted using the inverse document frequency (IDF) formula. The retrieved MESH headings carry the combined weight of the user input terms that map to them. The input terms, their variants and MESH headings identified as described above are used to search the database or ranked by weight and displayed to the user for selection. The first alternative is the application of automatic query expansion (AQE). In the second case only user selected terms are used in the retrieval process which is an example of interactive query expansion (IQE).

Term ranking mechanism used in CITE (Doszkocs, 1983) for IQE is based on the ranking method used in the Associative Interactive Dictionary (AID) reported by Doszkocs (1978). AID is a prototype system used for searching the Medline and Toxline databases. The system automatically extracts terms related to the user's query from the titles, abstracts and controlled vocabulary fields of retrieved documents. The strength of association between a term and a given query (a search term or a Boolean statement) are calculated by the 'relatedness value' (R), which is defined as:

$R=(O-E)/O$

where;

R: is the relatedness value representing the strength of association between a given term and a retrieved set of documents

O: is the frequency of occurrence of the term in the retrieved set

E: is the expected number of document occurrences of the term in the retrieved set

The expected number of document occurrences (E) is calculated by the following formula:

$E=(nT)/N$

where;

n: is the total number of documents in the retrieved set

T: is the number of documents in which the term occurs

N: is the total number of documents in the collection

When the observed frequency of occurrence of a term (O) is less then its statistically expected value (E), then the term is considered not to be semantically related to the user's query. The extracted terms are then ranked according their R values. The searcher may select terms from the ranked list for inclusion in her/his query.

The method described above is suitable in the context of Okapi searches, as terms are ranked according to their relatedness to the *whole* of the user query (see 7.3.1.1, 7.3.1.2, and 7.3.1.3). The above formula, therefore, is identified as a possible candidate to be used in the term selection process in the systems designed as part of the present project (7.3.1.3) and tested (8.3.1.4).

LEXIQUEST (described in Vickery & Vickery, 1993, pp. 127-128) is a prototype interface for online search. A particularly interesting feature of the system from the point of view of this project, is the search term 'normalization' mechanism whereby user input natural language queries are matched with controlled set of terms used in indexing the documents in its database. Although this system does not use a standard thesaurus, the index terms consist of compound as well as single terms. User input natural language query is analyzed by the system to identify individual content bearing words and phrases. These are then compared with the index terms. If identical index terms cannot be found in the system's lexicon for the extracted user terms, then term normalization procedure is applied. Normalization of terms make use of information about the co-occurrence of terms used to index documents. The formula used in the normalization process is given below:

$Z=(pAB)^2/(pA.pB)$

where;

Z: is the association measure between terms A and B

Pa: is the frequency of postings of the index term A in the database

Pb: is the frequency of postings of the index term B in the database

Pa.Pb: is the number of documents indexed by both A and B

Since the above formula is specifically used for normalizing the user input terms that match only partially with the compound index terms, it is identified as a suitable candidate for measuring the relatedness degree between user input terms and thesaurus terms in the systems designed in the present project (7.3.1.3) and tested (8.3.1.4). In the application of the above formula in the systems designed in this project, *Pa* is taken as the frequency of postings of the user input search terms (i.e. the number of documents in the retrieved set), *Pb* as the frequency of postings of the thesaurus term in question, and *Pa.Pb* as the number of documents indexed by both the user input search terms and the thesaurus terms (i.e. frequency of occurrence of the thesaurus term in the retrieved set).

Other systems make use of production-rules (i.e. if <condition> then <action> type formalism) similar to those used in expert systems to conduct a dialogue with the user in order to clarify the user input and a build a model of the search topic. An example of such a system, CIRCE, is described in Vickery & Vickery (1993, pp. 126-127). Users input queries in natural language. The system scans the user input to identify content words. These are then matched against terms in a thesaurus. The system then attempts to clarify the user input terms that partially match the thesaurus terms. The system generates questions, such as, "could you please clarify that input", "how is X related to Y", "does X mean the same as Y", by selecting the appropriate sentence template from a set of sixty.

IR-NLI (Brajnik et al., 1986), a front-end for online databases, incorporates a natural language module to extract useful search terms from user input natural language query description. The system engages in a dialogue with the user to construct a search strategy. The search tactics are stored in the system's knowledge base. These are derived from the work of Bates (1979). The system also incorporates a knowledge base that embody term relationships derived from a thesaurus which is used in search strategy formulation and query modification.

TomeSearcher (Vickery & Vickery, 1992) is another example of a front-end that uses a thesaurus and expert systems like rules for query formulation and modification. User input natural language text is analyzed to identify meaningful words and the relations between them. The user is consulted by the system to disambiguate multi-meaning words. User's query is then translated into its equivalent expression in Boolean logic. Domain specific knowledge stored in the form of a thesaurus is employed to formulate the user's query as a Boolean statement and modify the query when too few or too many documents are retrieved by the constructed search statement. The system also assists the user in selecting the appropriate databases for a given query.

IOTA (Chiaramella & Defude, 1987) combines expert systems and natural language processing techniques to aid the user in various stages of the retrieval process. The prototype system includes natural language processing of queries, user modelling, query understanding and re-formulation, management of full-text documents and relevance evaluation of answers. All knowledge used in the control of the system tasks is encoded as production-rules. Domain knowledge in the form of a thesaurus is used in query understanding and re-formulation processes. The system performs extensive syntactic analysis of users' natural language queries to map them onto index terms. The thesaurus is used in the mapping operation to help understand meanings of unknown query terms.

2.2.2 Spreading activation and thesaurus as a semantic network

The idea that human memory works by encoding the relationships between concepts based on the associations between them has been proposed by researchers working in cognitive psychology and artificial intelligence (AI) fields as early as sixties (Quillian, 1968; Minsky, 1968; Anderson & Bower, 1973). It was Quillian's (1968) semantic memory model that provided the first working computer model of human associative memory which has been since developed by others as a knowledge representation scheme in AI (Findler, 1979; Cohen & Feigenbaum, 1982). This model has been proved to be successful, particularly, in natural language processing and understanding applications (Schank, 1975; Miller, et al., 1990). In semiotics, Eco (1976) arrives a model of knowledge representation which is similar to that of Quillian's semantic memory model, from a different methodological and epistemological perspective (see 5.7.2)

Quillian's semantic memory network model consists of nodes representing concepts and links connecting nodes by any kind of (arbitrary) relationships. The resulting structure is usually known as a *semantic network*. In AI applications for NLP, nodes usually consists of verbs and other parts of speech which are linked according to certain kinds of relationships they acquire in natural language discourse. A Verb, for instance, is defined in terms of its relationship to an actor or a object (Cohen & Kjeldsen, 1987). HEARSAY-II (Erman et al., 1980) is an example of a speech understanding system based on a semantic network type of knowledge representation scheme and a black-board architecture which has been extensively adopted by researchers in knowledge engineering and expert systems fields (e.g. Nii, 1986). The system incorporates several independent knowledge sources based on the semantic network scheme. Each knowledge source represents expertise in a particular area of the speech recognition problem.

In information retrieval, Croft & Thompson (1987) adapted HEARSAY II's architecture in their I^{3}R system. The system has three distinct modules; interface manager, system experts, and knowledge-base. The interface manager provides a mechanism for communication between the users and the rest of the system. System experts are responsible for different aspects of the retrieval process, and communicate via a scheduler based on the black-board architecture of HEARSAY-II. Each system expert is made up of several production-rules. The following system experts are implemented:

- The user model builder: acquire information from the user regarding the goal of the search session, user's domain of interest and so on.

- The request model builder: obtains the user's query and represents it in a number of different forms, such as Boolean logic, and words or phrases associated with weights. This expert also obtains relevance judgements from the user on retrieved documents.

- The domain knowledge expert: search the knowledge base of the system to infer concepts that are related to those in the user's initial query. The found concepts are shown to the user for inclusion into the request model.

- The search controller: selects and executes search strategies implemented in the system. The search strategies are based on probabilistic and clustering models.

- The browsing expert: provides an alternative method to the keyword/concept based searching by maintaining a user-directed navigation activity in the knowledge base (i.e. browsing). The user can start browsing the knowledge base from a given document, author or index term and follow links to other items in the knowledge base.

- The explainer: provides explanations of the system taken actions in response to the user's requests. The explainer is based on similar techniques used in expert systems applications in AI.

The knowledge base of I^{3}R, represents documents, index terms, and other items as nodes in a semantic network. The knowledge base contains such items as, documents, terms and concepts used in indexing the documents, and document features, such as, author names, titles, and so on. There are several types of relationships in the knowledge base that link these items or nodes. Statistical document-document and term-term similarity measures for instance, are used to cluster documents and terms (this information is used in cluster searching). Bibliographic relationship links authors and documents, and semantic relationship is defined to connect synonymous terms.

Typical search session in I^{3}R starts by user typing a natural language query. Alternative methods of searching is also enabled in I^{3}R, the user for instance could start from a known relevant document. After the user and request models are built, the system instantiates one or the both formal search methods, namely, probabilistic and cluster searching. The user can activate the relevance feedback mechanism to refine the request model.

An important feature of the knowledge base in I^{3}R is the inference mechanism whereby, new concepts are derived from the user's initial query. The domain knowledge in terms of concepts are represented in I^{3}R similar to knowledge represented in a conventional thesaurus. A concept is represented as a *frame* in I^{3}R. Each frame consists of three parts: 1) the name of the concept, 2) information about how the concept can be recognized in text (documents), 3) the relationships between the concept and other concepts in the knowledge base.

A number of recognition rules in the form of "If <stem> then <concept> (degree of certainty)" are used to identify concepts in documents. Examples of such rules are:

"If <informat> then <information> (0.9)" or,

"If <information> and <retrieval> then Information_Retrieval"

The third part of a concept frame specifies the relationships of a concept to other concepts. The type of relationships used in I^{3}R are:

- Synonymy: represents the same concept

- Generalization: a narrower or broader concept

- Instantiation: the concept is an instance of another (e.g. Vax is an instance of Computer)

- Part-of: the concept is a component of another concept

- Cross-reference: this is an ambiguous relationship, which is used to link similar concepts, when the similarity cannot be defined by one of the above types of relationships

The domain knowledge in I^{3}R is constructed gradually through interaction with the users. This approach is not always the preferred strategy in other retrieval systems using semantic net type structures. Other systems described later in this section make use of preconstructed knowledge bases.

The recognition rules and relationships in the concept frames define an AND/OR tree structure of concepts. The system infers new concepts from user's query terms by traversing the AND/OR tree. For example, if Computer_Vision was inferred from the user input concepts of Computer and Vision, the next inference could be Pattern_Recognition which is linked to Computer_Vision by the cross-reference relationship in I^{3}R's knowledge base. The system then presents the inferred concepts to the user, and the user may accept the terms for inclusion in the new query.

The browsing expert in I^{3}R presents an alternative interaction mechanism to the users. The documents and other items in the knowledge base and the relationships between them are displaced graphically as a network of nodes and links. The user can select a node (e.g. a document) to examine its contents and then can follow any of the links that connects the document to other nodes in the network. The other nodes linked to the initial document may for instance, include the cited references in the initial document and the user could choose to examine the contents of any of the linked nodes and follow the links from them to navigate in the knowledge base in a similar fashion. This process is a semi-automatic version of the search procedure known as *spreading activation*, which is discussed in more detail below.

Another knowledge based information retrieval system, Metacat, which uses a semantic net type knowledge representation scheme and black-board architecture is reported in Chen (1992). Metacat incorporates eight knowledge sources, namely, the user model builder, the task model builder, known item instantiater, heuristic keyword searcher, thesaurus browser, online thesaurus, suffixing algorithm, stop word list. The general architecture of the system and many of the tasks performed by its knowledge sources are broadly similar to I^{3}R described in detail above, and therefore will not be discussed further here. Only, the online thesaurus and the thesaurus browser which implements a sort of automatic spreading activation method will be described in more detail as these are of particular interest to the present project.

The online thesaurus is a passive knowledge source that is activated by other procedural knowledge sources implemented in the system, such as the thesaurus browser. The online thesaurus consists of the portion of the Library of Congress Subject Headings (LCSH) and represented as semantic network in the system. The online thesaurus contains some 3500 terms in the areas of mathematics and computer science. Each of the terms has between a couple and a few hundred terms associated via one of the following relationship types: NT (used for a narrow term), BT (used for a broader term), RT (used for a related term), and USE (associates the synonymous terms). The description of the Inspec thesaurus used in the present project can be found in sections 7.1.1 and 7.1.2, which includes a more detailed description of the types of relationships found in thesauri.

If the known item and/or keyword search strategies invoked initially by the system do not yield satisfactory results, the thesaurus browser is invoked to search the online thesaurus to find terms that are related to the user's query. User input terms and other relevant terms identified by the system (e.g. terms from known relevant documents) are matched against the terms in the online thesaurus. Matching thesaurus terms are taken as *source nodes* by the thesaurus browser. The browser applies heuristic spreading activation process on the semantic network starting from the source terms. New nodes (terms) are activated by following the links leading from the source nodes, and a number of heuristics are applied to control the activation process which would otherwise grow exponentially and yield vast number of non-relevant new terms. The following heuristics are used by Metacat:

• The specific terms first heuristic: the system visits nodes that have fewer neighbours (links) before it visits nodes with more neighbours. This heuristic is based on the observation that in LCSH, terms with fewer links, in general, are more specific.

• The specific links first heuristic: the system attaches priority to the relationship types in the semantic net in the following order: NT, RT, and BT. Therefore, links of type NT is expanded before RT type links, which is expanded before BT type links. The heuristic is based on the idea that, the above order represents the specificity of relationships, and specific links lead to specific terms.

• The shorter distance heuristic: this heuristic states that the terms closer the source terms (i.e. separated by fewer nodes) are given priority. The rationale here is that, the further the distance between a source node and a term in the semantic network, the less relevant the term becomes.

• The two level expansion heuristic: states that, only terms that are two links away or less from a given source term are activated. This heuristic follows from the previous one, which suggests that the distance between two nodes is indicative of the semantic proximity of them. This heuristic is considered to be useful in finding terms that are semantically close to the source terms.

The above four heuristics are used by Metacat to compute a numeric figure called the 'relevance distance' which is indicative of the 'cost' associated with each path leading from the source terms to new terms. The paths are sorted according the relevance distance values associated with them and visited in the sorted order (i.e. paths with shortest relevance distance are expanded first). The spreading activation terminates when all source nodes have been connected or when all nodes that are two links away from the source terms have been activated. The expanded paths that are linked together are considered to address a similar underlying concept and the nodes on these paths are put in the same group. The result of the spreading activation process described above is usually a number of such term groups which are then ranked and presented to the user. The ranking is based on the number of source terms responsible for the generation of each group. Chen (1992, p. 307) suggests that, the four heuristics used are useful to control the spreading activation which otherwise costs a lot of computation time, and results in many irrelevant terms generated. Finally, user endorsed thesaurus terms are used to modify the original query.

Different researchers suggest different heuristics to be used in the spreading activation process. Shoval (1985) uses a 'metric of strength' based on the number of source terms involved in reaching a given term. The more the user terms involved in the expansion process, the more important the term reached in this way is assumed for a given query. Shoval also suggests that weighting of the links connecting the terms could be useful in ordering the terms. Kim & Kim (1990) reports a system that uses link weighting strategy which is described further below. Cohen & Kjeldsen (1987) reports an expert system, GRANT, which simulates the performance of a funding adviser. The system uses semantic network formalism to match researchers with funding agencies. GRANT's knowledge base consists of topics (words and phrases) representing research interests of funding agencies and researchers. Each topic corresponds to a node in a semantic network and linked to other topics via various types of relationships, including 'is-a', 'component-of', 'has-component', and 'object-of'. The system conducts a constrained spreading activation in the network starting from topics stated in a research proposal and spreading to nodes linked to the research topics until one or more topics representing research interests of agencies are activated. The system allows activation of up to four links from a research topic mentioned in the proposal. Since this is a weak constraint, additional constraints are applied to manage the activation process. One such constraint states that activation should terminate at nodes that have very high connectivity or fan-out. Examples of such nodes in GRANT's knowledge base include terms such as science, disease and person. Harter & Cheng (1996), propose the 'colinked descriptor hypothesis' as a basis for the spreading activation process in thesaurus. Harter & Cheng's method starts by selecting two terms from the ERIC thesaurus to describe the user's query. Terms that are linked to both of the original terms in the thesaurus are considered to be more likely to retrieve documents that will be judged relevant by the user.

The system reported by Kim & Kim (1990) uses a hierarchical concept graph (HCG) to represent the contents of a hierarchical thesaurus. Terms in a hierarchical thesaurus are connected by the 'is-a' relationship. HCG is a weighted hierarchical thesaurus where the nodes consist of thesaurus terms and edges (links) join the terms at lower branches of the hierarchy to those at the higher levels by the 'generalization' (BT) relationship. The edges are assigned weights which reflect the degree of generalizations between terms. Although, other types of relationships exist in conventional thesauri (such as, RT, and synonymy), in the reported system only the generalization relationship is used to simplify the inference mechanism. The system reported by Kim & Kim (1990) calculates the distance between any two nodes by summing the weights of edges along the path between the two terms. Queries are represented as Boolean statements, and both queries and documents are indexed by terms from the CRCS thesaurus (Computing Reviews Classification Structure). The system measures the conceptual distance between a query and a document by summing the edge weights of index terms as described above. The edge weights are manually assigned to the edges which is a very knowledge intensive and subjective work. For larger databases, it is clear that the task of manual assignment of weights becomes impractical and would almost definitely yield inconsistencies. Alternative methods of assigning link weights have been suggested (see further below).

Kim & Kim's work is based on the work of Rada and co-workers (Rada et al., 1989; Rada & Bicknell, 1989) who developed a document ranking method based on the assessment of conceptual distance between documents and queries using a hierarchical thesaurus (MeSH). Similar to the previous system described, Rada and colleagues represent both queries and documents as nodes in a hierarchical thesaurus, and matching consists of finding the shortest conceptual distance connecting the query and document terms. The conceptual distance between two terms are interpreted as the topological distance of the two terms in the hierarchical thesaurus. Documents connected to query terms with a shorter distance considered to be highly relevant for the query and ranked accordingly.

Other systems employ different strategies to assign weights to links which indicate the strength of association between two terms. In an experiment reported by Chen & Dhar (1991) weights of 3/9, 3/5, and 3/1 were assigned to relationship types NT, RT and BT, respectively. These weights represent the relative frequency of the links used by the searchers in an empirical study conducted by the above mentioned researchers which involved logging the behaviour of a number of users navigating a thesaurus. The rest of the system described in Chen & Dhar (1991) was similar to the Metacat system described earlier. In another experiment Chen and colleagues (Chen et al., 1993) asked the users to assign a numeric value of 0 to 10 to each of the relationship types NT, RT and BT, reflecting the users' preferences. These values are then used to modify the default weights of 3/9, 3/5, and 3/1 used in the previous experiment mentioned above (Chen & Dhar, 1991). Topic system (Chong, 1989) similarly assigns weights to links between topics (words and phrases) based on the user's assessment of the strength of association between them.

Alternatively, strength of association between two terms can be calculated by using the term co-occurrence data, that is, on the basis of their co-occurrence as document descriptors in the database. This approach is extensively used in automatic construction of thesaurus and similar knowledge structures and will be discussed in the next section.

2.2.3 Automatic/semi-automatic construction of thesauri and knowledge bases

It is generally agreed that, thesaurus construction is a very knowledge intensive and costly labour, especially in rapidly changing research areas, e.g. gnome research (Chen, et al., 1997). Researches have been looking for ways of constructing term relationship structures similar to conventional thesaurus automatically to overcome the difficulties involved in manual thesaurus construction and maintenance. Most of automatic thesaurus construction techniques involve statistical term co-occurrence algorithms (e.g. Salton, 1972; Crouch, 1990; Crouch & Yang, 1992; Chen & Lynch, 1992; Chen et al., 1995; 1997). Similarity coefficients between two terms are calculated based on their co-occurrence as index terms in a document collection, usually using a symmetric measure, such as the cosine function (Everitt, 1980).

Research in the area of term clustering over the past decades based on term co-occurrence analysis, however, suggest that limited benefit can be expected in using terms found by statistical techniques. Peat & Willet's (1991) research in particular suggest that similar terms identified by co-occurrence methods tend to have high frequency of occurrence in the database. Since high frequency terms poorly discriminate between relevant and non-relevant documents, they have severely limited effectiveness in query expansion. This observation supports the Sparck Jones' (1971) earlier findings that best retrieval results are obtained if only less frequent terms are clustered and used in the search process.

Chen and colleagues (Chen et al., 1995; 1997) approach the problem of automatic construction of term relationship structures by supplementing the co-occurrence analysis with a number of other methods. The approach proposed by Chen and fellow researchers involves three separate operations: document and object list collection, object filtering and automatic indexing, and co-occurrence analysis (Chen et al., 1997).

- Document and object list collection: The first step in any effort to build a thesaurus like structure involves collection of documents and other sources of vocabularies in a given domain. The purpose here is to identify domain-specific keywords which are useful in identification of important concepts in documents automatically.

- Object filtering and automatic indexing: Each document in the collection are matched with domain-specific keywords collected in the previous step. This process is referred to as object filtering. Since there may remain unidentified important concepts in the documents after the object filtering process, an automatic indexing procedure applied to extract further useful words and phrases. The procedure used is based on Salton's (1989) work on automatic indexing. First, individual words are identified, then words in the stop-word list are removed. This is followed by stemming operation on the words remain. Then, phrases are identified by combining only adjacent words.

- Co-occurrence analysis: The final step in this approach is to apply co-occurrence analysis for the terms identified in each document in the previous two steps. The term co-occurrence analysis is based on an asymmetric function developed by Chen & Lynch (1992).

The result of the above described procedure is a structure of term associations resembling the term relationships found in conventional thesauri. Other researchers tried different methods to construct thesaurus like knowledge structures.

Guntzer and colleagues (Guntzer et. al, 1989), used expert systems techniques to develop a system called TEGEN, which elicit term relationship knowledge interactively from end-users during online search sessions. The idea here is to tap the intelligence of a given user population exhibited during search sessions to construct a thesaurus that reflects expertise, interests and jargon of the population. TEGEN observes users' search behaviour, in particular, search statement formulation, to infer relationships among the search terms. Knowledge acquisition process in TEGEN consists of extracting relationships between search terms from the syntax of a search request by means of acquisition rules encoded in the form of production-rules. An example rule is given below (Guntzer et. al, 1989, p. 268):

if (a) two or more search terms x_i in a search request X are combined by OR
and
(b) x_i occur as keywords
and
(c) the search request X is combined in further search request Y by AND or AND NOT with further search terms y_j
then
produce similarity relationships among all x_i

The results of above type analysis are recorded as intermediary results by the system and shown to the searcher for verification.

Another approach to automatic thesaurus construction is described in Raghavan & Jung (1989). The method described here is called 'pseudo-thesaurus' construction and based on machine learning (in particular neural network) techniques. The idea is to construct term relationship structures for specific user groups by replacing the term frequency and co-occurrence information used in term clustering by user relevance judgements. In the approach developed by Raghavan & Jung both positive and negative term-term relationships are considered.

2.3 Semiotics and Information Retrieval

The objective of this section is to summarize the research on the relationship between semiotics (including cognate fields, such as philosophy of language, and speech acts theories) and information retrieval.

It seems, there is a general consensus in information science community that information retrieval is about *communication* of information among humans. It is sufficient to look at the proceedings of the first conference on Conceptions of Library and Information Science to appreciate this fact (Vakkari & Cronin, 1992). However, there has been relatively little research done so far to analyze the information retrieval dilemma as a human communication problem. The general objective of the research described in this dissertation is to do just that, that is, to analyze information retrieval process as a human communication phenomenon and develop a retrieval method based on this analysis. Since semiotics is a powerful conceptual tool to analyze human communication and sign systems (see chapter 3), it has been employed in the present project as a general methodological framework.

There have been relatively little but nevertheless significant amount of work done in the past which attempt to develop an understanding of the information retrieval situation from the perspective of semiotics and speech acts or language games theories. Blair's research is perhaps the best known example of such an effort.

In 'Language and Representation in Information Retrieval', Blair (1990) attempts to develop a coherent and exhaustive analysis of the retrieval problem mainly from the perspective of speech acts theories of Austin (1962), Searle (1969), and language games theory of Wittgenstein (1953), although semiotics of Saussure, Eco and others as well as Pierce's pragmatics have also been made use in his analyses. I do not intend to depict a rich picture of his work which covers a large area of activities related to information retrieval, but merely summarize his main arguments and conclusions here.

Blair (1990; 1992) analyzes information retrieval problem from a pragmatic view of the nature of human communication and language. From the point of view of language games and speech acts theories (see also 3.9 and 5.5.5) participating in an communicational act (utterance of verbal and non-verbal signs) is like participating in a 'game', say, chess. From this perspective, there are types of linguistic utterances or acts which are not primarily governed by conditions of 'truth'. In such cases, language is used to *do* things or perform certain action rather than denote or describe things in the real world. Austin (1962) calls these types of linguistic acts as 'performatives' or 'illocutionary acts'. Some examples of illocutionary acts are (in Blair, 1992):

- I'll pay you right back

- I name the ship the 'Norton Sound'
- Finish the report before tomorrow's meeting
- Bill's a better worker than Bob

In all of the above cases, each utterance configures the sender (or addresser), the addressee and the referent of the utterance in a certain way, in short, it makes something *happen*. According to Blair, same sort of analysis can be extended to document retrieval systems to understand how certain implicit and explicit conventions function in document retrieval. Blair gives the case of citations at the end of scholarly articles as an example of an illocutionary act in the context of document retrieval. According Blair, the inclusion of a citation in the bibliography of the article is an example of a certain type of illocutionary act, namely, 'assertive declaration', where the author "... *declares* that the citation(s) are part of the bibliography of the article, and asserts that the articles they refer to are relevant to the citing article" (Blair 1992, p. 202).

Blair, in a manner similar to the above described case of citations, analyzes various components of the document retrieval process using the speech acts and language games theories. His main conclusion is that, retrieval systems should be designed in such a way that the various speech acts embedded in documents, indexing languages and retrieval mechanisms are made explicit, so that the users and designers of such systems share similar contexts and participate in the same language game. The following excerpt from Blair (1992, p. 206) summarizes his main points regarding design of retrieval systems:

"In the first place, the language of document retrieval, like ordinary language must have its meaning grounded in activities. Consequently, there won't be one way of describing a document, but a variety ways, each based on the activity that uses the document in question. Thus, information retrieval systems are activity specific. They, like their language, are dependent on the activities that they serve. The role of the indexer or the designer of indexing algorithms is to relate the usage of the terms which represents the documents to the usage of those words in the activities that employ those documents. As a result, the study of information retrieval can be thought of as the study of information *in context*. We cannot separate the design of the retrieval systems from the activities in which they are embedded ..."

A similar view of information retrieval is described in Brier (1996). Brier, employing the second order cybernetics of Maturana, Bateson, Luhmann, and others and pragmatics of Peirce and Wittgenstein as well as speech acts theories, attempts, like Blair, to connect individual users' information retrieval behaviour to the larger context of social practices.

Another relevant strand of research can be found in the work of Andersen (1990; 1986). Andersen analyzes computers as 'sign systems' or 'media'. He makes use of conceptual tools developed by several important semiotic schools, in particular, post-Saussurean semiotics of Hjelmslev and Eco (cf. 3.4 and 3.10). His main concern is to analyze computer interfaces as media through which humans communicate. Consequently, Andersen's work concentrates on the investigation of computer-based signs and involves pragmatic, rhetorical and aesthetic levels of analysis. He attempts to relate design of computer interfaces to specific work languages and pragmatics of performing tasks in specific work contexts. Although his work does not directly address the problems involved in information retrieval, he applies semiotic concepts to work situations which include those associated with document retrieval activities (Andersen, 1986).

Warner (1990; 1994), using semiotic concepts aims to bring documents and computers within a single analytic category of 'text'. He attempts to formulate a unifying principle based on the understanding that both computer programs and documents are products of human semiotic

- I name the ship the 'Norton Sound'
- Finish the report before tomorrow's meeting
- Bill's a better worker than Bob

In all of the above cases, each utterance configures the sender (or addresser), the addressee and the referent of the utterance in a certain way, in short, it makes something *happen*. According to Blair, same sort of analysis can be extended to document retrieval systems to understand how certain implicit and explicit conventions function in document retrieval. Blair gives the case of citations at the end of scholarly articles as an example of an illocutionary act in the context of document retrieval. According Blair, the inclusion of a citation in the bibliography of the article is an example of a certain type of illocutionary act, namely, 'assertive declaration', where the author "... *declares* that the citation(s) are part of the bibliography of the article, and asserts that the articles they refer to are relevant to the citing article" (Blair 1992, p. 202).

Blair, in a manner similar to the above described case of citations, analyzes various components of the document retrieval process using the speech acts and language games theories. His main conclusion is that, retrieval systems should be designed in such a way that the various speech acts embedded in documents, indexing languages and retrieval mechanisms are made explicit, so that the users and designers of such systems share similar contexts and participate in the same language game. The following excerpt from Blair (1992, p. 206) summarizes his main points regarding design of retrieval systems:

"In the first place, the language of document retrieval, like ordinary language must have its meaning grounded in activities. Consequently, there won't be one way of describing a document, but a variety ways, each based on the activity that uses the document in question. Thus, information retrieval systems are activity specific. They, like their language, are dependent on the activities that they serve. The role of the indexer or the designer of indexing algorithms is to relate the usage of the terms which represents the documents to the usage of those words in the activities that employ those documents. As a result, the study of information retrieval can be thought of as the study of information *in context*. We cannot separate the design of the retrieval systems from the activities in which they are embedded ..."

A similar view of information retrieval is described in Brier (1996). Brier, employing the second order cybernetics of Maturana, Bateson, Luhmann, and others and pragmatics of Peirce and Wittgenstein as well as speech acts theories, attempts, like Blair, to connect individual users' information retrieval behaviour to the larger context of social practices.

Another relevant strand of research can be found in the work of Andersen (1990; 1986). Andersen analyzes computers as 'sign systems' or 'media'. He makes use of conceptual tools developed by several important semiotic schools, in particular, post-Saussurean semiotics of Hjelmslev and Eco (cf. 3.4 and 3.10). His main concern is to analyze computer interfaces as media through which humans communicate. Consequently, Andersen's work concentrates on the investigation of computer-based signs and involves pragmatic, rhetorical and aesthetic levels of analysis. He attempts to relate design of computer interfaces to specific work languages and pragmatics of performing tasks in specific work contexts. Although his work does not directly address the problems involved in information retrieval, he applies semiotic concepts to work situations which include those associated with document retrieval activities (Andersen, 1986).

Warner (1990; 1994), using semiotic concepts aims to bring documents and computers within a single analytic category of 'text'. He attempts to formulate a unifying principle based on the understanding that both computer programs and documents are products of human semiotic faculty and can treated as different varieties or genres of what is commonly referred as text.

Although does not include direct examination of information retrieval systems, Liebenau and Backhouse's (1990) analyses of communication and information in organizational settings, and in the context of Management Information Systems have important consequences for the design of retrieval systems as noted by Brier (1996, p. 335-336). Semiotic analyses applied by Liebenau and Backhouse at the levels of empirics, syntactics, semantics, and pragmatics underline the importance of the social setting and the context of activity in communication of information and creation of meaning.

Although does not directly apply a semiotic framework, Hjørland's (Hjørland 1992; 1997; 1998; Hjørland & Albrechtsen, 1995) work explicate the importance of 'knowledge domains', 'social practices' and 'discourse communities' in understanding the various aspects of the document retrieval process and in this regard related to the semiotic analyses of document retrieval described above. One of the important results of Hjørland's (1992) work is the characterization of the 'subjects' of documents as their 'epistemological potentials'. This important formulation presents a solution to the problems caused by the mentalistic approach of the cognitive viewpoint in information science which is criticised for being idealistic and reductionist, among others, by Frohmann (1990). More recently, Hjørland (1997; 1998) presents analyses of the document retrieval activity at various levels with a view to determine the usefulness of different 'subject access points' in databases. This work can be said to pave the way for a 'database semantics'. It is impossible to present a detailed and accurate description of the work carried by Hjørland which covers a wide spectrum of activities related to document retrieval here, however I would like to note that, there are similarities at various levels between the approach developed in the present dissertation and the work represented by Hjørland, some of which are pointed out in later parts of this dissertation.

Frohmann (1990; 1992; 1994) presents a critique of the cognitive view of information retrieval which he characterize as 'mentalistic'. Although, his analysis is not directly based on semiotics, he applies a methodology which is related to pragmatics, and especially to language games theory of Wittgenstein.

Finally, the present author gives a somewhat detailed account of document retrieval systems from a semiotic view point (Karamuftuoglu, 1996; 1997). A more recent work (Karamuftuoglu, in press) applies the semiotic viewpoint to explicate the challenges faced by retrieval systems design practice and theory in relation to emerging network centric computing applications and attempts to reformulate the basic concepts of 'relevance' and 'user' in light of the developments in network based information systems and services.

Chapter 3
A Survey of Semiotics

In this chapter some of the basic concepts of *semiotics* are discussed. In the following survey, concepts and principles related to both of the major schools of semiotics, namely, that of (post-) Saussurean *structuralist* tradition and Peircian semiotic tradition, are introduced. Only those topics that will prove to be useful for the arguments of the rest of this dissertation are covered. In this capacity, this chapter does not claim any comprehensive discussion of semiotic concepts and categories. The following basic introduction to semiotics will serve as a background to the 'semiotic' analysis of IRS in chapters 4, 5, and 6.

3.1 Preliminaries

Semiotics is the *science of signs*. This definition requires clarification of what is meant by both 'science' and 'sign'.

Scientific status of various disciplines has been object of some very well known controversies in the history of Science. In particular, scientificity of some 'human' and 'social' sciences has been questioned. See for example Popper (in Miller, 1983, pp. 119-130) for a discussion and definition of what constitutes a 'scientific' discipline. Popper's criterion to differentiate the scientific from the non-scientific or pre-scientific is to look in the theory for the possibility of being refuted by empirical evidence or testing. Thus any theory, field, or discipline which closes itself to some form of possible falsification or refutation, according to Popper, is not a science. However this is not a clear cut criterion as it is not always possible without more elaboration and 'philosophical' discussion to say what constitutes a refutable theory. A detailed discussion of the above mentioned questions are of course best left outside of the limits of this dissertation.

Without getting my self into more trouble, I can safely change my definition to: *Semiotics is a discipline which studies sign systems*, thus avoiding the problem of what constitutes a scientific theory.

The object of study of semiotics is 'sign' (cf. 3.2), and it is concerned with everything which can be taken as a sign. Therefore, semiotics studies everything and anything (sic) that manifest the semiotic correlation, its domain is the whole of the human culture, an immense range of objects and events that can be taken as a sign from a particular point of interest (Eco, 1976, pp. 6-7). As Bense puts it: "A sign is anything that is declared a sign, and nothing but that is declared a sign" (in Nake, 1994, p. 194). One can call this sort of endeavour as 'general semiotics' and it is a philosophical and not a scientific discipline because of this generality. It studies the whole of human signifying systems -- i.e. all existing languages (Eco, 1984, p. 12).

However, semiotics also concerns with specific systems of signification, and in this case one can talk about 'specific semiotics'. A specific semiotics aims at being the grammar of a particular sign system. Being a well defined field with specific goals and limitations and specific epistemological problems, specific semiotics is a scientific field, in so far as other 'human' sciences are, according to Eco (ibid, p. 5).

Alternatively, one can instead adopt a more sophisticated classification by borrowing Hjelmslev's proposal according to which, there are a 'scientific semiotic' and a 'non scientific semiotic', both studied by 'metasemiotic'; a 'semiology' as a 'metasemiotic' studying a 'nonsemiotic semiotic', whose terminology is studied by 'metasemiology' (ibid, p.4).

Eco (1976, pp. 9-14) lists following research areas as belonging to 'semiotic field': zoosemiotics, olfactory signs, tactile communication, codes of taste, paralinguistics, medical semiotics, kinesics and proxemics, musical codes, formalized languages, written languages, natural languages, visual communication, systems of objects, plot structure, text theory, cultural codes, aesthetic texts, mass communication, and rhetoric.

I would like to add to this list semiotic approaches to design of computer systems and HCI, as most notably exemplified by Andersen (1990): "Semiotics is the science of signs and their life in society. Its subject is all kind of signs: verbal language, pictures, literature, motion pictures, theatre, body language. *Computer semiotics* is a branch of semiotics that studies the special nature of computer-based signs and how they function in use" [my *emphasis*] (p. 3). One can note as other major attempts to establish a 'semiotic' foundation for 'informatics', Gorn (in Machlup, 1983, pp. 121-140) and Slamecka and Pearson (in Weiss, 1977, pp. 105-128). Nadin (in Deely, 1985, pp. 463-470), Nake (1994), Desouza (1993) discuss the semiotics of computer interface design. Warner (1990; 1991) explores the relation between *Information Science* and *Semiotics*, for whom Information Science has affinities with Semiotics in that, they both have interest in the products of human semiotic faculty, such as documents, texts, words, etc. (1990, p. 17). Blair (1992; 1990) examines the relation between information retrieval and the *speech acts* theories of Austin, Searle Griece, and *language games* theory of Wittgenstein (cf. 3.9) in detail.

In this connexion, I would like to note that, one of the main arguments of this dissertation is: 'IR research as carried out in the context of Information Science would benefit from a semiotic perspective'.

3.2 Sign

As mentioned above, the object of semiotics is sign, and it is necessary to elaborate on it. A sign is a correlation between a signifier and a signified, an *expression* and its *content*[1][2]. This is the formulation given by the pioneer of the field Ferdinard de Saussure (1974, p. 66) : "The linguistic sign unites, not a thing and a name but a concept and a sound-image. ... The linguistic sign is then a two-sided psychological entity that can be represented by the drawing:

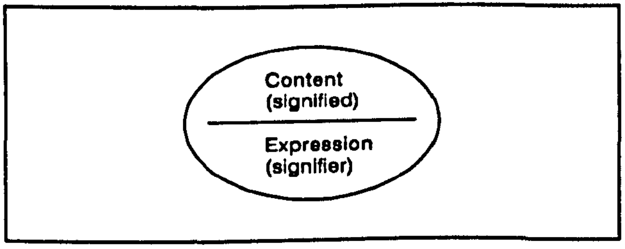


Figure 3.1: Sign

[1] This is a postulate of semiotics (Eco, 1976).

[2] Hjelmslev calls these *expression plane* and *content plane*, respectively (Barthes, 1967, p. 49). See section 3.5.10 for more discussion of Hjelmslev's formalisation of the 'sign'.

The two elements are intimately united and each reveals the other".

Saussurian definition of sign is therefore, a double faced entity, a 'Janus', two faces of the sign, signifier and signified is inseparable from one another and they make together the two faces of the same and the single entity like two faces of a coin. The important point in this description is the correlation between the two faces of the sign. Therefore, a sign is the totality of the *signifier* and the *signifieds* in the process of a semiotic correlation.

For the sake of terminological clarity and to emphasise this functional aspect of sign it is best to follow Hjelmslev's example and to call this correlation a 'sign-function' and each of its members as 'functives' of this relation. One can therefore safely say that a sign-function arises when an expression (sign-vehicle) is correlated to a content (concept) (Andersen, 1990, p. 69).

Pierce[3] articulate the 'sign' in terms of the following definition (in Eco, 1976, p. 15) : "(a sign is) something which stands to somebody for something in some respect or some capacity". It is already clear in this definition that, a *sign-vehicle* which is present is standing for something else (signified) which is absent and the process of signification which leads from signifier to signified is an inferential one, i.e. imputed. Saussure's term for the 'mediated' or 'contractual' relation between the 'signifier' and the 'signified' is: arbitrary: "... Not only the two domains are linked by the linguistic fact shapeless and confused but the choice of a given slice of sound to name a given idea is completely *arbitrary*" [my *emphasis*] (Saussure, 1974, p. 113). The correlation between the signifier and the signified is arbitrary, in the sense that the association between them "... is the outcome of a collective training" [my *emphasis*] (Barthes, 1967, p. 50). One can therefore, say that the link between the signifier and the signified is unmotivated[4] (see 3.5.3 for more discussion of this).

3.3 Antecedents

Since semiotics covers such a vast area of study including philosophy, metaphysics, anthropology, sociology, logic, medicine, and more; it is beyond the scope of this dissertation to do a thorough survey of its antecedents. The following is a brief discussion of some of its origins.

The earliest sources of contemporary semiotics can be found in the ancient medicine which concerned with the sensible indications of changes in the condition of the human body (Sebeok, 1976, pp. 3-4).

Semeion -from sema 'sign' and semeiotikos 'observant of sign'- appeared as a technical-philosophical term with Parmenides and Hippocrates in the fifth century B.C. It is often found as a synonym of 'tekmerion' (proof, clue, symptom). A first distinction between the two terms did not appear until Aristotle's (384 - 322 B.C.) rhetoric (Eco, 1984, p. 26).

However, it was the Stoics who gave the term a broader meaning, and it became a basic division of philosophy, including logic and the theory of knowledge. In the Middle Ages, a

[3] Whose contribution to semiotics is discussed in more detail in sections; 3.5.1, 3.5.2, 3.5.3, and 3.5.4.

[4] Except for the cases of 'onomatopoeia' and 'propositional' signs (Barthes, 1967, pp. 50-51).

comprehensive theory of signs, embracing grammar, logic, and rhetoric, was elaborated by a number of scholars.

This line of investigation is carried forward by Leibnitz (1646-1716), who studied the syntactical features of sign-structures. His ideas have been adopted and developed by symbolic logicians and others such as Boole, Carnap, Frege, Husserl, Russel, Tarski, Whitehead, etc. The most important exponent of this tradition is however, C. S. Peirce (1839-1914) whose contribution is examined in more detail below. In contrast to the above mentioned syntactics oriented line of investigation, there was a empirically oriented line of semiotic enquiry, which investigated the semantic dimension of signs as exemplified by Francis Bacon, Berkeley, Hobbes, Hume and most notably by Locke (1632-1704) who introduced the Stoic term 'semeiotike' into English philosophical discourse in his essay 'Concerning Humane Understanding' (Sebeok, 1976, p. 4; Kristeva, 1989, pp. 295-296).

At the beginning of this century, semiotics took a whole new direction with Saussure (1974) and his programme of natural languages oriented 'semiology', which will be discussed in little more detail below.

3.4 Semiotics and/or Semiology

As we have seen above it is possible to identify two major streams of orientation in today's semiotics. First one is Locke-Peirce-Morris tradition ultimately going back to the Stoics, which is prevalent especially in North America. It is a logic-philosophy oriented tradition, which took a behaviouristic inclination especially with Morris in the middle of this century. Its most important representative today is perhaps Sebeok. In fact neither Peirce nor Morris had ever used the form 'Semiotics', their favoured form was 'Semiotic'. "I am as far as I know, a pioneer, or rather a backwoodsman in the work of clearing and opening up what I call *semiotic*, that is the doctrine of the essential nature and fundamental varieties of semiosis" (Peirce quoted in Eco, 1976, p. 15). However somehow 'Semiotics' had eventually become the preferred term in English language in the second half of this century (see Sebeok, 1976, for an interesting account of this, pp. 48-52).

The second stream of contemporary semiotics is linguistics oriented, originating mainly from the work of Saussure (1857-1913). It has its antecedents in the ancient medical tradition (Sebeok, 1976, p. 53) and has found repercussions in the (post-) structuralist continental European linguistics and philosophy (such as; Benveniste, Jakobson, Hjelmslev, Eco, Barthes, Derrida, Baudrillard, Lyotard, Kristeva, Foucault).

Saussure envisioned a general science of signs, roughly at the same time, yet independently from Peirce, which he named as Semiology: "A science that studies the life of signs within society is conceivable; it would be a part of social psychology and consequently of general psychology; I shall call it semiology. Semiology would show what constitutes signs, what laws govern them. ... Linguistics is only a part of the general science of semiology; the laws discovered by semiology will be applicable to linguistics, and the later will circumscribe a well-defined area within the mass of anthropological facts" (Saussure, 1974, p. 16).

Semiology has ever since widely spread throughout French scientific, linguistic discourse (Sebeok, 1976, p. 55) although it is not the only form. Kristeva (see for example, the reference list in Sebeok, 1976, p. 226) for instance, who is an important exponent of French linguistic-semiotic discourse, prefers 'semiotique' in contrast to Saussure's 'semiology'.

In his influential volume 'Elements of Semiology', Barthes (1967, p. 11) adopts Saussure's term semiology, however, for him inverting Saussure's hierarchy, it is semiology that has to be a part of linguistics. Barthes posits natural languages in a privileged position in relation to semiology because of the pervasiveness of language in every sphere of the social realm: "... it is far from certain that in the social life of today there are to be found any extensive systems of signs outside human language ... every semiological system has its linguistic admixture" (ibid, pp. 9-10). Barthes hopes that: "By this inversion we may expect to bring to light the unity of research being done in anthropology, sociology, psycho-analysis and stylistics round the concept of signification" (ibid, p. 11).

This anthropocentric position however ignores the semiotic studies which deals with non-human species (Sebeok, 1976, p. 165). It has also been criticised as being verbocentric by others, since substantial part of semiotics concerns with non-verbal signs, which can not be readily translated into verbal units (see for instance Eco, 1976, pp. 172-174).

Barthes' position however stresses the strong link between *linguistics* and *semiotics* which is already evident in Saussure and certainly the link between two disciplines is much more than being trivial: "Saussure's originality consisted of the recognition of the vital importance for linguistics of a comparative analysis and classification of different sign systems ..." (Sebeok, 1976, p. 10).

In summary it can be concluded that, the word semiotics has connections with philosophical discourse in the Locke-Peirce-Morris mould and its the preferred form especially in American semiotics whereas, semiology originating from Saussure has linguistics connections, and tends to be more abundant in French texts (Warner, 1990, p. 29; Sebeok, 1976, p. 56).

Although the differences associated with the terms semiotics and semiology are not mutually exclusive but simple contrasting tendencies (Warner, 1990, p. 29), there are a number of reasons here for opting for one rather than the other. I opt for semiotics in this dissertation, mainly because I am concerned with all sorts of expression-vehicles that can be taken as signs, which may or may not resemble the linguistic sign. The term semiotics implies broader scope for signification, which is not restricted to linguistic items such as words and morphemes. Although, we are obviously dealing with linguistic entities in information retrieval, their structure might well not be identical to the linguistic 'lexeme'[5]. The importance of this will become clearer when sign in IR context is examined thoroughly in chapters 4 and 5.

3.5 Foundations of Modern Semiotics

Sebeok (1976, p. 181) uses the metaphor of a tripod to illustrate the foundation of modern semiotics. At one side there is Saussure and his linguistics, at the other side there is Peirce who stands as 'the heir of the whole philosophical analysis of signs', and the third somehow uneven leg is that of medicine, the ancestral figure being Hippocrates.

Most contemporary semioticians (see e.g. Eco, 1976; Kristeva, 1989; Sebeok, 1976) agree that modern semiotics has been developed almost simultaneously, yet independently, by Saussure in Europe and Peirce in the United States. The following sections aim to introduce some basic ideas of modern semiotics which owe much to Saussure and Peirce, in somewhat more detail.

[5] For definitions of morpheme and lexeme, see section 3.6.1.

3.5.1 The Semiotic Triad

From the preceding definition of sign (3.2) we have arrived the following structure so far: *there is a physical entity, a sign-vehicle, which refers to something other than itself and this relation is recognized by someone.* Peirce in the U.S. and Ogden and Richards in Britain arrived a similar model to the above description of sign (Fiske, 1982, p. 45).

In Peirce's words: "A sign stands for something to the idea it produces, or modifies....That for which it stands is called its object; that which it conveys, its meaning; and the idea to which it gives rise, its interpretant" (in Eco, 1979 p. 69).

In Peirce's theory, the sign is therefore, a triadic relation, between an object its representamen, and the interpretant (Kristeva, 1989, p. 13). Ogden and Richards represent this as :

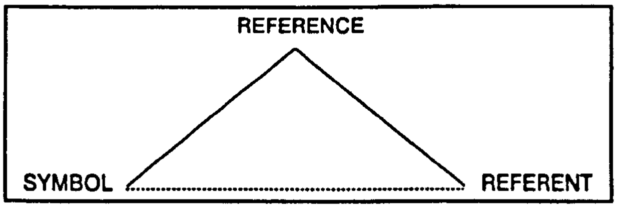


Figure 3.2: Symbol/Reference/Referent (in Eco, 1976, p. 59)

which corresponds to Peirce's triad (Eco, 1976, p.59):

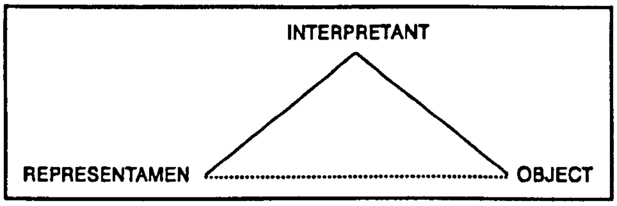


Figure 3.3: Representamen/Interpretant/Object (in Eco, 1976, p. 59)

At the first glance this triadic relation (as formulated variously by Peirce, and Ogden and Richards) can be read as "the word signifies -- the thing -- by means of mediating concepts" (Lyons, 1977, p. 96). I will go on to argue in chapter 4 that, this is in fact the general understanding of 'signification' in Information Science.

This interpretation of the triadic relation is rather superficial however, and certainly not correct in the case of Peirce (see Eco, 1976, pp. 58-60 and 68-69).

It is worthwhile to note here that, the other main formulation of sign (i.e. Saussure's), concerns only with the left side of the above triad. Saussurian sign is a relation between a sign-vehicle and its content, therefore, the third element of the Peircian triad, object has been left out by Saussure only to assume a secondary role (Eco, 1976, p. 60). In Saussure's description the sign relates to reality only through the concepts of the users of that sign (Fiske, 1982, p. 44). However, when we analyze the elements of Peirce's triad in more detail in the following section, we will see that Peirce's object is also subsumed to secondary importance in the process of signification, i.e. semiosis.

3.5.2 Unlimited Semiosis

Semiosis is the Peirce's term for the process of signification: "By semiosis I mean an action, an influence, which is, or involves, a cooperation of three subjects, such as a sign, its object and its interpretant, this tri-relative influence not being in any way resolvable into actions between pairs" (in Eco, 1976, p. 15). As Posner (1992, p. 49) puts it succinctly, semiosis is "The process in which something functions as a sign, that is a process in which some A, interprets some B as representing C".

The most subtle element of this relation is the interpretant. "The Interpretant is not the interpreter.... The interpretant is that which guarantees the validity of the sign, even in the absence of the interpreter" (Eco, 1976, p. 68).

For Peirce, the interpretant is another sign translating and explaining the first one (ibid, p. 15). In order to establish what the interpretant of a sign is, it is necessary to name it by another sign, which in turn have another interpretant to be named by another sign and so on ad infinitum (ibid, pp. 68-69). "... Peirce underlines that the interpretant is also a sign. That is, the interpretation is contextually, socially, and historically determined, and is therefore constantly developing" (Brier, 1992, p. 102).

In Peirce's own words: "The object of a representation can be nothing but a representation of which the first representation is the interpretant. But an endless series of representations, each representing the one behind it, may be conceived to have an absolute object as its limit. ... Now the Sign and the explanation together make up another Sign, and the explanation will be a Sign, it will probably require an additional explanation, which taken together with the already enlarged Sign will make up a still larger Sign; and proceeding in the same way, we shall, or should ultimately reach a Sign of itself, containing its own explanation and those of its significant parts; and according to this explanation each such part has some other part as its Object " (in Eco, 1976, p. 69). This final object can not be any object but the entire *semantic field* (3.7.1), thus concludes Eco (ibid).

The unlimited semiosis as described above, results in infinite regression of meaning, which leaves a little place for external referential object (cf. 3.5.14).

3.5.3 Symbols, Icons, Indices

This is another trichotomy originating from the work of Peirce. Peirce constructed a classification scheme which divides signs into 10 classes with further sub-divisions resulting ultimately in sixty-six varieties (Lyons, 1977, p. 100). This was the most comprehensive and subtle effort in the history of semiotics (Sebeok, 1976, p. 120). However, his classification were based on intersecting criteria (Lyons, 1977, p. 100) and it is generally suffice to consider the three main categories of signs:

The icon refers to the object it is representing through its similarity with it. An example of an icon is a design of a tree that represents a real tree by resembling it (Kristeva, 1989, p. 13). A sign is said to be iconic when there is a topological similarity between a signifier and its denotata (referent) (cf. 4.2.1, 4.2.2) (Sebeok, 1976, p. 43).

The index, on the other hand, does not necessarily resemble the object that it is referring, but is affected by it, therefore has something common with it; e.g. smoke is an index of fire

(Kristeva, 1989, p. 13). A sign is said to be indexic in so far as its signifier is contiguous with its signified, or is a sample of it (Sebeok, 1976, p. 43).

The symbol in Peirce's sense, refers to an object that it designates by convention, agreement or rule, that is by the intermediary of an idea (Kristeva, 1989, p 13). It is arbitrary in this sense (cf. 3.2). This corresponds to linguistic 'sign' of Saussure[6]. A sign without either similarity or contiguity, but only with a conventional link between its signifier and its denotata with an intensional class for its designatum (signified) is called a symbol (Sebeok, 1976, p. 43).

Although, this is proved to be a considerably useful classification, it has its own problems which become important when we are interested in analyzing IRS in terms of sign systems. Much more detailed discussion of this is given in 4.2.

In addition to these three fundamental categories of sign, three others identified by Peirce worth to be mentioned here: Signal, Symptom and Name.

When a sign-vehicle mechanically or conventionally triggers some action on the part of the receiver. it is said to function as a signal. Examples of signals are: the exclamation 'Go!' or, alternatively, the discharge of pistol to start a footrace (a conventional releaser v. a mechanical trigger). The term is particularly useful in animal communication studies (Sebeok, 1976, pp. 121-124).

Eco in (1976, pp. 20-21), uses a slightly different terminology. For the above definition, the term Eco uses is stimuli, i.e. Sebeok's signal becomes stimuli in his terminology (cf. programmed stimuli in 5.3.3). Signals for Eco are units of transmission which can be computed quantitatively irrespective of their possible meaning. At this point semiotics confronts with its lower threshold.

A Compulsive, automatic, nonarbitrary sign, such that the signifier is coupled with the signified in the manner of a natural link, is called a symptom. A syndrome is a rule-governed configuration of symptoms with a stable designatum (signified). Both terms have strong, but not exclusive medical connotations (Sebeok, 1976, pp. 124-128).

A sign which has an extensional class (cf. 3.9) for its designatum (signified) is called a name. Thus individuals denoted by the proper name 'Veronica', have no common property attributed to them save the fact that they all respond to 'Veronica' (ibid, pp. 138-140).

3.5.4 Syntactics, Semantics, Pragmatics

This trichotomy has been formulated by Morris, goes back ultimately to Peirce (Lyons, 1977, p. 114).

Morris took Peirce's thought and developed it into a behaviouristic semiotics which was consistent with the intellectual climate of America in the '30s (Sless, 1986, p. 144).

For Morris "Semiosis (or sign process) is regarded as five-term relation -*v*, *w*, x, *y*, *z*- in which v sets up in w the disposition to react in a certain kind of way, *x*, to a certain kind of object *y*

[6] Oddly enough, symbol in Saussure's (1974) terminology is a sign with a motivated link to its signified, corresponding to what Peirce calls icon and index.

(not then acting as a stimulus), under certain conditions *z*. The *v*'s, the cases where this relation obtains, are signs, the *w*'s the interpreters, the *x*'s are interpretants, the *y*'s are significations, and the *z*'s are the contexts in which the signs occur" (in Kristeva, 1989, p. 298).

Syntactics as generally understood, studies the combinatorial rules of signs, in Morris's words "...syntactics deals with combination of signs without regard for their specific significations or their relation to the behaviour in which they occur" (in Lyons, 1977, p. 115). Posner (1992, p. 40) differentiates three senses of syntactics: *i)* the study of formal aspects of signs, *ii)* the study of the relations of signs to other signs, *iii)* the study of the way in which signs of various classes are combined to form complex signs.

Semantics deals with the relation between the sign and what it signifies or mean. According to Morris "...semantics deals with the signification of signs in all modes of signifying" (in Lyons, 1977, p. 115). Semantics according to Morris takes into account signs and objects, but not interpreters (Sebeok, 1976, p. 14). In Posner it is defined as: "The study of the conditions an entity must fulfil so that can be represented by signs for interpreters in semiosis" (1992, p. 49).

Pragmatics is the epistemologically uppermost layer of this trichotomy. It is "The study of the conditions an entity must fulfil to be able to interpret signs as representing meaning in semiosis" (ibid). Morris defines it as "... that portion of semiotic which deals with the origin, uses, and effects of signs within the behaviour in which they occur" (in Lyons, 1977, p. 115). Thus, according to Morris pragmatics takes into account all three factors, namely, sign, its object and its interpreters (Sebeok, 1976, p. 14).

From Morris' behaviour oriented point of view, pragmatics is the most important concern of semiotics. It is claimed by behaviourial and logic oriented semioticians that, pragmatic relationship presupposes the semantic and the syntactic; the semantic presupposes only the syntactic; and the syntactic presupposes neither (see e.g. Doede, 1972, pp. 40-42). However, Sebeok (1976, pp. 14-15), gives counter evidence to this from zoosemiotics, i.e. animal sign use.

3.5.5 Langue and Parole

One of the major contributions of Saussure to linguistics/semiotics, is his proposal to study language as a 'system'. This makes possible to study language in its totality, as a whole: system with its internal structure made of *differences*. The 'value' of each element of the system emanates from its relation to the other members of the system: "... language is a system of independent terms in which the value of each term results solely from the simultaneous presence of the others" (Saussure, 1974, pp. 114-115). "A linguistic system is a series of differences of sound combined with a series of differences of ideas; but the pairing of a certain number of acoustical signs with as many cuts made from the mass of thought engenders a system of values; and this system serves as the effective link between the phonic and psychological elements within each sign. ... Although both the signified and signifier are purely differential and negative when considered separately, their combination is a positive fact; it is even the sole type of fact that language has, for maintaining the parallelism between the two classes of differences is the distinctive function of the linguistic institution" (ibid, pp. 120-121). This is the **paradigmatic** (cf. 3.5.8) aspect of language.

The above formulation of *language as a system*, is called 'la langue' in Saussure's terminology. It is a social system -- i.e. a system of *convention* -- that is been there before the individual, therefore outside direct control of individual members of the society: "... the social side of language, outside the individual who can never create or modify it by himself; it exists only by virtue of a sort of social contract signed by the members of a community" (Saussure, 1974, p. 15).

This formulation effectively divides language into two halves; language as: 'language-system' (la langue) and 'speech' (la parole).

Language as a system has an autonomous existence, independent of "... what is spoken by any subject" (Kristeva, 1989, p. 9), and made up of signs which stands in opposition to one another (i.e. signs exist as differences in relation to each other). Language-system is the underlying schema, which makes possible the manifestation of speech. Speech also called 'language-usage' or 'language-behaviour' (e.g. Lyons, 1977, p. 26) is always individual, realized through the speaking subject by means of the underlying schema that is the language-system.

Therefore, according to Saussure, language need to be studied in two distinct parts: as a system which is a social phenomenon, independent of the individual user of that language, and as a language-usage or speech which is individual and manifested through the underlying system. Speech or language-use is the syntagmatic (cf. 3.5.8) aspect of language.

The two parts, in fact are inseparable from each other. La langue is a prerequisite for speech, and similarly la langue can not exist as an abstract entity without any usage or speech.

3.5.6 Synchrony/Diachrony

Examination of language as a system requires to view it in its totality with its given structure, and precise operational rules, complete at any given moment. This is called synchronic study of language, which is for its most part ahistorical and universalist. "By the synchronic analysis of a language is meant the investigation of the language as it is, or was, at a certain time" (Lyons, 1977, p. 243).

On the other hand language is subject to changes in time, being transformed in different eras and among different people. Diachronic linguistics studies evolution of language through history. "By the diachronic analysis of a language is to be understood the study of changes in the language between two given points in time. If we apply strictly the distinction of the diachronic and the synchronic, we will say that the notion of one language (e.g. English) existing over the centuries ... is fallacious" (ibid).

Saussure's innovation rests on his careful separation of the study of language as a system of terms organized as opposition, from the study of changes between terms that are succeeding one another in time: "What diachronic linguistics studies is not relations between co-existing terms of a language-state but relations between successive terms that are substituted for each other in time" (Saussure, 1974, p. 140).

Linguistic studies in the nineteen century is marked with 'comparative philology' and 'historical linguistics'. The main concern of this linguistics is the comparison of different 'language-families' and their 'evolution' throughout history with an overall aim of finding the 'common origin' of languages and especially that of 'European languages' (Kristeva, 1989, pp. 191-216; Saussure, 1974, pp. 1-5). "Languages were grouped into families by deriving each member from an initial source" (Kristeva, 1989, p. 195).

The epistemological breakthrough brought afore by Saussure, marked the birth of a new linguistics which is descriptive and studies language proper (la langue) as a system, replacing the nineteenth century's historical (diachronic) linguistics.

It is obvious that one would like to re-unite this two apparently mutually exclusive points of view on language. This has been exactly the aim of many post-Saussurian linguists.

For those who emphasise (such as Prague and Copenhagen Linguistic circles) the synchronic nature of language, history is implicated in synchrony (Kristeva, 1989, pp. 223-237). According to the solution found to this problem by post-Saussurian structuralist linguistics, each of distinct language-systems that exist at different points in history can be studied synchronically and independently from each other, and diachronic linguistics can investigate how an earlier system was transformed into a later system (Lyons, 1977, p. 243). Therefore, diachronic linguistics presupposes and dependent upon synchronic study of language-systems, as autonomous, independent systems made up of interrelated, differential elements. "The temporal model proposed by Saussure is that of a series of complete systems succeeding each other in time; that language is for him a perpetual present, with all the possibilities of meaning implicit in its every moment" (Jameson, 1972, p. 6).

To conclude this section, synchronic/diachronic division which has brought about by Saussure, made possible to analyze language in two spheres; system (langue) and speech (parole). The separation of langue from the rest of 'language mass' made possible to define an unified and classifiable 'object' for modern linguistics (Kristeva, 1989, p. 9).

3.5.7 Paradigmatic and Syntagmatic Relations

There are two basic types of relations that units of a language-system contract; syntagmatic and paradigmatic.

The syntagmatic relations are those contracted when an element of a language-system combines with other elements of the same level. This is also called the horizontal axis (Hammarstrom, 1976, p. 5; Lyons, 1977, p. 240). The parole, or speech that we have identified above as distinct from and outside the langue is mainly concerned with this axis. This is the axis which concerns the individual during the act of speech or communication.

Having said that, the rules governing the combinatorial possibilities of any syntagmatic chain is dependent upon the underlying language-system or schema (Lyons, 1977, p. 241).

The syntagmatic relations are realized at different levels of the system. For example, in English 'i' is syntagmatically related to both 'p' and 't' in the written word-form 'pit'. In a larger syntagmatic chain, 'the old man', the lexeme 'old' syntagmatically related with the definitive article 'the' and the noun 'man' (ibid, pp. 240-241).

The paradigmatic (called associative in Saussure) relations are those, between an element found in some particular context and another element which could have been substituted, in place of the one under consideration in the same syntagm. This is called the 'vertical axis' (Hammarstrom, 1976, p. 5; Lyons, 1977, p. 241).

The paradigmatic relations as with the syntagmatics ones, hold at different levels of the language-system; e.g. the letters 'i', 'e', 'a' are paradigmatically related in the word-forms such as 'pit', 'pet', 'pat', etc. Similarly, 'old' is paradigmatically related to 'young', 'tall', in the expressions like 'the old man', 'the young man', 'the tall man', etc., just as 'man', 'woman' and 'dog' are intersubstitutable for one another in expressions like 'the old man', 'the old woman', 'the old dog' (Lyons, 1977, p. 241).

The paradigmatic and the syntagmatic relations in language facilitate the selection (the choice axis or the paradigmatic axis) and combination (the combination axis or the syntagmatic axis) of linguistic units at every level (Berry, 1977; Lyons, 1977).

3.5.8 Communication and Signification

Semiotics deals with two distinct but complementary phenomena; *communication* and *signification*. The distinction between these two phenomena need to be carefully demarcated.

The most common, although somehow simplistic definition of communication, as usually given both by semioticians and communication scholars alike, is: communication is the passage (transmission) of message from a source to a destination (Cherry, 1968; Fiske, 1982; Kristeva, 1989; Hervey, 1982), which can be illustrated by the following simple schema (in Kristeva, 1989, p. 8):

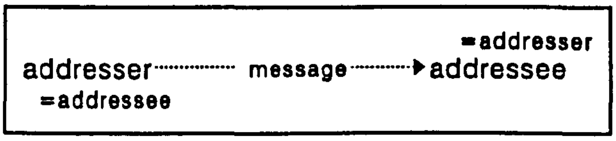


Figure 3.4: A Simple Communication Model

The most important aspect of human communication is the social context in which it takes place. This implies the existence of shared or common values, or conventions in human communication (Cherry, 1968, p. 4). Human communication can therefore be defined as 'social interaction through messages (Fiske, 1982, p. 3).

It is possible to categorize the current perspectives on the communication phenomenon roughly into two distinct schools, which are named as the 'process school' and the 'semiotic school' by Fiske (ibid, pp. 2-3).

The process school views communication as the *transmission of messages*. It is concerned with the accuracy and efficiency of the transmission, and its effects (behaviourial or cognitive) on the addressee. It analyzes the communication process in terms of the channel of communication, intention of the sender/receiver and the content of the transmitted message (Fiske, 1982, p. 3; Halloran, 1983, p. 160).

The semiotic school view communication as *the production and exchange of meaning*. It is mainly concerned with the role that texts play in the process of production and exchange of meanings or signification, and the social dimension within which the interaction between messages, texts and people takes place (Fiske, 1982, pp. 2-3).

Eco (1976, p. 8) defines communicative process "... as the passage of a *signal* (not necessarily a sign) from a source (through a transmitter, along a channel) to a destination" [my *emphasis*] (ibid). In this case of course one can not speak of signification but passage of information[7],[8] (Eco, 1976, p. 8; Moles, 1966, p. 19). Only when the destination is a human being (or other intelligent biological being or mechanical device)[9] one can speak of signification and therefore passage of a sign or message, on condition that the signal is not merely a stimulus but arouses an interpretative response in the addressee. The interpretive response, i.e. signification, is made possible by the existence of a culturally defined system or schema (cf. 3.5.5) that underlies the communication process. According to Eco therefore, the demarcation should be made between 'semiotics of communication' and 'semiotics of signification' (cf. 3.5.9) (Eco, 1976, p. 8).

To conclude this section; when we speak of communication, it is usually understood as the process of exchange of messages by means of the existing shared social values or conventions (Sebeok, 1976, p.1). This is made possible because of the underlying system of signification (Eco, 1976 p. 8).

3.5.9 Code/S-code

The underlying system is usually referred as *code* in semiotics and communication studies (e.g. Cherry, 1968; Fiske, 1982; Eco, 1976): "A code is a system of signification, insofar as it couples present entities with absent units" (Eco, 1976, p. 8). A code is therefore, a semiotic system that establishes correlation between an sign-vehicle and a content-unit (cf. 3.2).

One should note here that natural language or langue is probably the most important of all codes, albeit a code among many others (and not all of them are verbal) which constitute the complex web of human communication. Not only there are multitude of codes at play at any instance of communication (Eco, 1976) but 'langue', as conceived by Saussurian (structuralist) linguistics, itself is, but one of plurality of overlapping signifying systems (Kristeva, 1989, p. 296).

Similar to the langue/parole (system/speech) relation described in the previous section, any communication process or act presupposes and is dependent upon the underlying code. On the contrary a code has an autonomous existence, independent of any potential communication process that might take place by using that code. The addressee's cognitive or behaviourial response is not necessary for the definition of the code (cf. 3.5.9). However, their interrelation is a strong one and they are always intertwined in actual 'cultural processes' (Eco, 1976, pp. 8-9).

Therefore we can speak of two objects of study for semiotics, that of semiotics of communication which studies communication process as sign usage and production and semiotics of signification which deals with the structure of codes or underlying system which makes

[7] See section 3.5.11 for a distinction between 'signification' (or 'meaning') and 'information'.

[8] It is perhaps useful to say communicative processes involves 'signals' rather than 'messages', in this sense, in order to be able to classify 'machine-to-machine' transmissions as a communication process as well.

[9] Assuming that there exists such a device or machine.

possible the communication process[10] (ibid).

Eco distinguishes further between codes and s-codes. S-codes are pure oppositional structures without any correlational or communicational purpose. They are made up of finite number of elements that are governed by combinatorial rules, that generate finite or infinite chains of these elements (ibid, p. 38). S-code or code as a system constitute only one of the planes of the signification process or semiosis (cf. 3.7.1). In this sense, it is similar to the idea of paradigm (3.5.7) or language-system (la langue) of Saussure (cf. 3.5.5). When an s-code is correlated with another, a correlational function that we have called code above emerges. This is the conception of code as an equivalence structure between the two planes of signification[11]. Eco (1984, pp. 164-188) calls this, the weak sense of the term code, and discusses alternative conceptions of it that he calls the strong sense of code (cf. 4.4.3).

3.5.10 Expression/Content, Form/Substance

Saussure, as discussed in 3.5.5, conceives linguistic system (la langue) as pairing of a series of differences on the plane of sounds with a series of differences on the plane of ideas or thought: "The linguistic fact can therefore be pictured in its totality -i.e. language- as a series of contiguous subdivisions marked off on both the indefinite plane of jumbled ideas (A) and the equally vague plane of sounds (B). ... The characteristic role of language with respect to the thought is not to create a material phonic means for expressing ideas but to serve as a link between thought and sound, under conditions that of necessity bring about the reciprocal delimitation of units. ... language works out its units while taking shape between the two shapeless mass" (Saussure, 1974, p. 112). For Saussure, thus, languages result from imposition of structure or form on the shapeless mass of sound and thought, or in other words, on the sound and thought substance (Lyons, 1977 p. 240).

Hjelmslev, who was one of the most important proponents of the post-Saussurean structuralist tradition, formalizes this aspect of language by dividing each of the planes of the sign-function, namely the 'expression' and 'content' planes, into purport or matter, substance and form. According to this model, expression-form transforms the purport or matter the expression is made of[12], into expression-substance, i.e. material occurrences (tokens). Similarly, content-form transforms the content-purport[13] into content-substance, i.e. positive (emotional, ideological, notional, etc.) meaning (Barthes, 1967 pp. 39-41; Kristeva, 1989, pp. 235-236; Eco, 1976, pp. 51-54). The form, i.e invariant features of a sign-substance (expression or content substance) is determined by a procedure called the 'commutation test' (see 3.6.1).

[10] See the introductory paragraphs to chapter 5 for the differences between the two objects of studies; i.e. a 'theory of sign production' and a 'theory of codes'.

[11] See the next section for a discussion of 'the planes of signification'.

[12] Such as; sounds, graphic material on paper, luminous material on computer screen, etc.

[13] That is, unorganized mass of thought, emotion, so on (Lyons, 1968, p. 56).

3.5.11 Information, Meaning

It is necessary to clarify the relation of information and meaning in order to introduce the methodological distinctions imposed by semiotics.

It is not the object of this section to discuss various interpretations of the term 'information' in various contexts, but to make the distinctions between 'semantic information' and measure of the 'amount of information' formal.

From semiotics point of view information and signification need to separated carefully. Signification as noted in 3.5.10 is the function of the code which correlates a sign-vehicle with a content unit. In this sense meaning is the effect of this semiotic correlation; i.e. the content unit becoming meaning of the signifier. Any transmission of information which acts merely as stimuli on the receiver/destination and does not cause a semiotic process i.e. semiosis, is said to be devoid of meaning.

When we speak of information in semiotics, it is not the semantic information (i.e. signification) but 'syntactic-information' or 'signal-information' (cf. 3.5.8). This distinction between semantic-information and signal-information as quantified by the 'mathematical theory of communication' or 'information theory' of Shannon and Weaver (1949) has been made by some authors, variously calling it; signal-information, syntactic information, amount of information, or simply information (Bar-Hillel, 1964; Cherry, 1968; Eco, 1976; Lyons, 1977; Moles, 1966; Wilden, 1977). Information as measured by information theory is therefore devoid of meaning/signification. It is a measure of the complexity of the structure of a message (Moles, 1966, pp. 53-54).

It may be useful from semiotics' perspective to further distinguish between 'information at the source' and 'information transmitted' (Eco, 1976, pp.40-46).

Information at the source denotes the information at one's disposal when a code is selected to compose a message, thus it is a statistical property of the source, designating the possible selections that can be made from a source[14], as studied by the mathematical theory of communication. Information (i.e. signal or syntactic information) in this sense has a content (i.e. meaning), only by virtue of its *potential* for making selections (Cherry, 1968, p. 171).

Information transmitted is the amount of information that is actually transmitted when a selection made from the all possible (equi-probable) information at one's disposal at the source.

3.5.12 Message, Text

What is transmitted in a communication process is usually referred as message, both in the information theory and semiotics.

A message is often defined as a finite ordered set of elements of perception drawn from a repertoire of signs (e.g. alphabet) and assembled in a structure with an intention to signify or communicate (Cherry, 1968, pp. 171, 307; Moles, 1966, p. 9).

[14]'Source' can be thought as a repertoire of symbols for the sake of simplicity.

From the communication engineering point of view (cf. 3.5.8), which sees communication as a process of transmission of messages, *intention* is the crucial aspect of any message. "The message is what the sender puts into it by whatever means" (Fiske, 1982, p. 3). Therefore, it is mainly interested in efficient and accurate coding/decoding and transmission of messages.

Semiotics, on the other hand, emphasize in the communication process the signification aspect, and therefore the generation of meaning. As we have seen above (3.5.8, 3.5.9) signification or semiosis as it is usually called, involves correlating a content-unit (or an aggregate of content units) to a sign-vehicle by means of cultural conventions. Therefore, content of a sign is a product of nothing else than the cultural space it is produced in[15]. It is a cultural unit (cf. 3.7) (Eco, 1976, pp. 62-66).

This shifts the focus from the message itself to how to read or decode the message within the social context of the communication act. Every signification is a result of a culturally established code (cf. 3.5.8). However, as noted earlier, there are a number of codes in play at any given moment of the signification process. Therefore, what is ordinarily called a message is in fact a multilevelled *discourse* which is a product of multiple codes superimposing many levels of meaning (signification) upon one another (Eco, 1976, pp. 57-58). Reading is the process of discovering meaning*s* that occur when the addressee interacts or negotiates with the text. The negotiation takes place when the reader brings aspects of her/his cultural background to interpret the signs which make up the text (Fiske, 1982, p. 3-4).

When we say message, from semiotics perspective, it is therefore actually a text, whose content is a (multilevelled) discourse (Eco, 1976, pp. 57-58). This is a very important distinction, which illuminates the difference in the conception of the communication process between the semiotic and process (the communication theory) oriented schools of view.

To explicate the *multilevelled* structure of signification more clearly, we need to examine the concepts of 'denotation' and 'connotation', which is the subject of the next section.

3.5.13 Denotation, Connotation

When we analyzed the structure of sign in 3.2., we said that, sign is made up of two inseparable parts; the signifier and the signified, and signification or meaning results from their correlation. This is the first level of signification and usually referred in semiotics as denotation (Barthes, 1967, p. 89).

The above relation can be formulated as: **E R C** where; **E** is the expression plane (signifier), **C** is the content plane (signified) and **R** is the semiotic relation between the two planes (ibid).

There are systems however, such that, the relation **E R C** becomes the expression plane of another similar system, which can be schematically shown as: **(ERC) R C** . This relation is called connotation in semiotics. A *connoted* system is therefore, a system, whose plane of expression is itself constituted by a signifying system (ibid, pp. 89-90).

[15] This is true even for the so called 'iconic' signs, that are assumed to be similar, or related by some natural bond, to their referent (cf. 4.2.1 and 4.2.2).

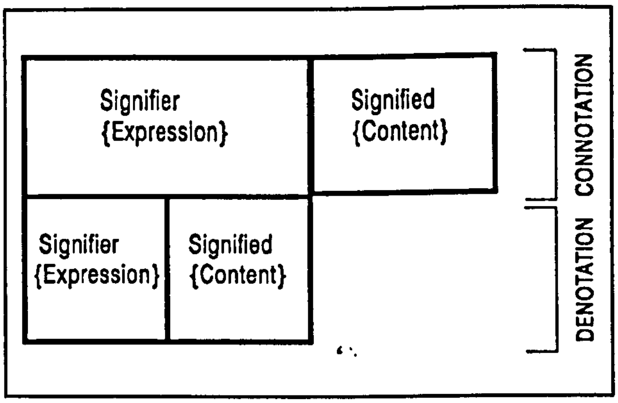


Figure 3.5: Denotation/Connotation

Although some authors (e.g. Bangura, 1994, p. 160; Lyons, 1977, pp. 206-215) define denotation in terms of the relation of the sign to the external objects (denotata) or the class of external objects (denotatum) that it refers to, Eco (1976, pp. 58-59) demonstrates that this is not necessary. For the methodological reasons that will become clear in the course of this dissertation (see especially section 4.2.3), this should actually abandoned altogether. Furthermore, when the intension/extension dichotomy is discussed in the next section, it will be shown that the idea of referent does not lead to useful criteria in analyzing various types of signs.

The difference between the two codes, the denotative and the connotative, is not a difference between a univocal and vague signification, or between referential and emotional, etc (as suggested by, e.g. Bangura, 1994). The difference is that, the connotative code relies on a primary one (the denotative) (Barthes, 1967, p. 91). The choice of a primary code is to do with the coding convention, although it is usually true (but not necessary) that connotations are usually less stable than denotations (Eco, 1976, pp. 55-56).

3.5.14 Intension, Extension

When Peirce's typology of signs in 3.5.3 is discussed, we have defined the category of name in terms of it having an external class for its designatum.

The extension of a term is defined in propositional calculus, as the class of things to which it is correctly applied. Therefore, extensional definition of a word, say 'dog', can be done by listing its all (sic) members (Lyons, 1977, p. 158).

The intension[16] of a term, on the other hand, is the set of essential properties which determines the applicability of the term. The intensional definition thus, identifies a class on the basis of some property which all its members have in common (ibid, pp. 158-159).

Logical semantics and more specifically, *propositional calculus* deals with the 'truth-values' of propositions. In standard propositional calculus, the propositions may have one of the two possible truth-values; that of true or false (ibid, pp. 141-142). When the truth-value of a proposition is considered in propositional calculus, it is determined by the extension of its terms. Similarly the truth-value of a complex proposition is determined by the truth-value of its constituent propositions (ibid, p. 160).

[16] Note the spelling.

Therefore, a sentence such as "all dogs have four legs" is considered to be true in propositional calculus ('extensional semantics' or 'theory of extension', or 'theory of reference' as sometimes called), if and only if all dogs really have four legs. Theory of extensions, is thus, concerned with the actual state of the world that a proposition corresponds/refers to (Eco, 1976, p. 62).

The meaning of a term or a statement (proposition) as understood by extensional semantics should not therefore be confused with that of semiotics. Semiotics (or more specifically theory of codes, cf. 3.5.9) is only interested with the meaning insofar as, there is a semiotic function which correlates an expression unit with a content unit. As it would be readily recalled from section 3.5.12 that, a content unit is nothing more than a cultural convention, a cultural unit[17], therefore from semiotics point of view, meaning of a term or a sentence is independent of the truth-value of the corresponding proposition, (i.e. the extension of the term/proposition) and the corresponding state of the world (Eco, 1976, pp. 62-66). Thus, Eco repeatedly says: "Every time there is a lie there is signification. Every time there is signification there is the possibility of using it in order to lie" (ibid, p. 50). This means that if there is a possibility of semiotic correlation between a given sign-vehicle and a given 'content unit' in a given code, there is signification i.e. semiosis, irrespective of the truth-value of the corresponding proposition or state of the world.

As mentioned in section 3.5.13, the sort of semiotics we have adopted, for the methodological reasons that will become clear later, leaves a little, if any, place for external objects or states. Semiotics, as adopted for the purpose of this dissertation, opts for the intensional definition of meaning, as understood in the sense discussed in the preceding paragraph.

For the time being, the following would illustrate clearly, why any semiotic approach which aims to be inclusive of enormous variety of items that can be taken as a sign, could not fruitfully adopt an extensionally oriented point of view:

There are many terms in any language, which do not apply to any object or thing that exists in the universe, such as 'centaur' and 'unicorn', both of which can be said to have the same extension, which is the empty class (Lyons, 1977, p. 159). These are *culturally* defined entities, which are communicated by means of other signs (cf. the concept of semiosis in section 3.5.2). In addition to the terms whose extension is null, there is the whole category of syncategorematic terms, such as 'to the', 'of', and 'nevertheless', which do not have any extension or referent (Eco, 1976, pp. 66-67).

In any case it should be extremely difficult if not impossible to define any term extensionally. For example, the extension of the term 'dog' is all existing dogs. "But all existing dogs is not an object which can be perceived by the senses. It is a set, a class, a logical entity ... which moreover is only a cultural convention" (ibid, p. 66).

All these considerations forces one to abandon the 'referent' oriented semiotics (extensional) for a non-referential semiotics (intensional). This was the main orientation of the semiotics of both Saussure and Peirce as noted in 3.5.2.

[17] For the structure of 'cultural units', see 3.7.

3.6 Units of Semiotic Systems

In some semiotic systems, such as the human language, it is possible to identify distinct, invariant features of the system which make up the elements of the signification process (i.e. 'functives' of the expression and content planes in Hjelmslev's terminology).

As we shall see later (3.6.2) not all semiotic systems can be analyzed, in a similar fashion, into more analytic elements.

3.6.1 Units of the Verbal Language

It has been noted in 3.5.5 that, the human language as conceived by Saussure and his followers (most notably Jakobson and the Prague school, Hjelmslev with his Glossematics and the Copenhagen School) is made of series of 'differences'. The value of the each element of this system is purely the function of its position in relation to the other elements of the same system (Malmberg, 1967, p. 13).

The elements of both the expression and the content planes are built up from smaller discrete elements which can be discovered by segmenting a text or a utterance progressively into smaller portions and applying the 'commutation test'.

The commutation test is used in various modern linguistics schools as a tool for establishing invariant elements (cf. 3.5.10) of a linguistic system. It works from larger units progressively into smaller ones, until non-meaning pertaining units are discovered that are not signs, therefore the correlation between content and expression planes ceases to exist (Malmberg, 1967, pp. 14-15). "The commutation test permits an exhaustive functional (linguistic) segmentation of a soundwave[18], i.e. it gives us the smallest possible independent linguistic units on the expression level[19]. Any commutable segment which cannot be split up by a new commutation into still smaller elements is a minimum unit" (ibid, p. 72).

To illustrate how the commutation test works, consider the following example: In the morphemes 'bill', and 'pill', which are member of a much larger series (paradigm) such as: *bill, pill, kill, mill, hill, till, chill,* etc., only one unit varies: /b/, /p/. By substituting /p/ for /b/ one gets 'pill' which has a different meaning than 'bill', i.e., a corresponding change occurs in the content plane. Therefore, it is concluded that these two units /p/ and /b/ commute. They are distinctive (invariant) elements of the expression plane. It can be shown by similar analysis that, they can not be segmented into yet smaller elements, therefore, one can conclude that minimal units of the expression system has been identified (Gleason, 1961, pp. 15-16). These minimal units of the expression system is called *phonemes*.

3.6.1.1 Phonemes

Human languages are shown to organize the selection that they make of the available sound differences in human speech into a limited number of recurrent distinctive units, called phonemes (Robins, 1980, p. 101).

[18] Can be equally well applied to other 'matter' (purport).

[19] Can be equally well applied to the 'content plane'.

"The phoneme is the minimum[20] feature of the expression system of a spoken language, by which one thing that may be said is distinguished from any other thing which might have been said" (Gleason, 1961, p.16). In written language the corresponding units are 'letters' (Lyons, 1977, p. 232). The total number of phonemes in any language is limited, and on average between 30-40 (Aitchison, 1972, p.17).

The phonemes are in themselves meaningless, but they combine into larger grammatical units called 'morphemes', which are meaningful. This is called the 'second articulation' (Malmberg, 1963, p. 15).

3.6.1.2 Morphemes

Morphemes are generally short sequences of phonemes. They are the smallest units of language which are endowed with meaning (Gleason, 1961, p. 51). They can similarly be identified by pulling out progressively smaller portions of a text and applying the commutation test (Hockett, 1969, pp. 123-124). The number of morphemes in any language is enormous compared the number of phonemes. One could expect a competent speaker of a language to make use of in the order of $10^{3.}$ morphemes (Aitchison, 1972, p. 17).

It is sufficient for most purposes to assume that words are made up of morphemes. Therefore, it is usually accepted that the English word 'unacceptable' is composed of three morphemes, 'un', 'accept', and 'able' (Lyons, 1968, p. 181).

Morphemes are combined into broader syntagms to form larger units, such as, 'the foxhunter' which is composed of the following morphemes 'the + fox + hunt + er'. This is called the 'first articulation' (Malmberg, 1967, p. 15).

3.6.1.3 Lexemes

In the preceding section we had made reference to the grammatical form word as a complex meaningful unit composed of simpler units, namely morphemes.

It is important to distinguish between a phonological word, such as the one orthographically represented as 'sing' and its inflextional forms; 'sang', 'sung' and 'singing'. To mark this distinction a more abstract category of 'lexeme' is used. Thus, in modern linguistics term word refers to phonological (or orthographic) forms such as 'sing' and 'sang', whereas lexeme denotes abstract units which occur in different inflexional forms according to syntactic rules. Thus, the orthographic word cut represents three different inflexional forms, all of which is of the lexeme cut (Lyons, 1968, p. 197).

[20] Some schools of linguistics, such as the Prague circle, and especially Jakobson, maintains that *phonemes* can further be analyzed into yet smaller components called 'distinctive features' (Lyons, 1977, p. 232). These are universal, and number a dozen in all languages (Kristeva, 1989, p. 228).

3.6.1.4 Sememes

All the units we have discussed in the preceding paragraphs belong to the expression plane of the language system.

As noted in the preceding sections 'Structural Linguistics' of Saussure supposes that the language system maintains two parallel and inseparable planes, that of Expression and Content which alone makes possible the functioning of language. Hjelmslev's Glossematics (cf. 3.5.10) in particular, asserts that expression and content planes are structured in "quite analogues fashions" (in Kristeva, 1989, p. 236). Structural Sẹmantics attempts to establish the units of the contents plane, as Structural Linguistics establishes the units of the expression plane (Lyons, 1977, pp. 231-269; Eco, 1976, pp. 75-76).

The corresponding element in the content plane is the sememe (Chao, 1968, pp. 73-73), which denotes some semantic component of the expression units. It has been variously called 'sememe', 'semantic marker', 'semantic category', etc. (Lyons, 1968, pp. 470-471). Various different definitions of sememe is given by different schools. It is formulated to correspond to the lexeme in Greimas' structural semantics (Schleifer, 1987, p. 71); for Bloomfield (1961, p. 162), it is the meaning of a morpheme. In the rest of this dissertation sememe is used as the meaning of a sign-vehicle (i.e. the content unit that is correlated to a sign-vehicle), regardless of whether the sign-vehicle in question is a verbal one or not.

Hjelmslev uses the term 'figurae' to denote the minimal non-meaning bearing elements (i.e. elements of the second articulation) of both the content and the expression planes. Therefore, phonemes can be said to be expression figurae (Malmberg, 1967, p. 15).

The structure of the *semantic space* will be analyzed in more detail when the concept of 'semantic field' is introduced in the next section (3.7).

3.6.2 Units of Non-Verbal Languages

It is worthwhile to ask the question, whether there are functionally contrasting units in other sorts of sign-systems, similar to the units found in phonology (phoneme) and verbal language in general (morpheme, lexeme, sememe, etc.).

One can look at, for example, graphic representation of objects, that is, iconic signs, and try to establish the pertinent features. The question to be addressed is: are there distinct units in the sign-system (i.e. the graphic or iconic representation) which are culturally coded and analyzable into more analytic units? This is the alternative way of saying, are there iconic phonemes or sentences? If there are indeed culturally coded pertinent units in the graphic substance, are they subject to double articulation (cf. 3.6.1.1 and 3.6.1.2), similar to the verbal language?

The answer to this, is no in the case of iconic signs. Even if one can identify distinct units in a graphic-language they are not durable, i.e. same unit represent totally different thing in a different context. More accurately, iconic figurae seem not to correspond to linguistic phonemes because they do not have positional and oppositional value (Eco, 1976, pp. 213-216).

However, there are some non-verbal systems which do have similar structure to the verbal language, and subject to single or double articulation such as, naval signs (second articulation only), traffic signs (fist articulation only), deaf alphabet, morse code, tonal music (both first and second articulations), so on. One can even think of codes with three articulations. For example, the cinematographic language offers this possibility (ibid, pp. 232-234).

Indeed several new terms invented and proposed to describe the units of various non-verbal semiotic systems thought to be analogous to the functional aspect of the linguistic phoneme, such as; 'behavioureme' (general idea of behaviourial pertinent unit), 'kineme' (minimal unit of facial expression or body gesture), so on (Crystal, 1971, pp. 183-186).

3.7 The Semantic System

When intensional and extensional approaches to meaning is discussed in 3.5.14, it is noted that, meaning cannot be defined in terms of the external referent satisfactorily, as there are various categories of linguistic units that do not have a referent. It is also said in the same section that, meaning of a sign-vehicle is a cultural-unit.

Peirce's conception of the signification process, in terms of unlimited semiosis (3.5.2) makes perfectly clear that each sign is interpreted, explained, by other signs, functioning as its interpretant. Signs are communicated to us by means of other signs that defines, interprets the initial sign. In this sense the meaning of a sign-vehicle is a cultural-unit, communicated by means of other expression-vehicles: "... *cultural units are physically within our grasp.* They are the signs that social life has put at our disposal: images interpreting books, appropriate responses interpreting ambiguous questions, words interpreting definitions and vice versa" (Eco, 1976, p. 71). This explains perfectly well, how an extensionally null sign, such as "unicorn" conveys a meaning, although obviously there is no corresponding referent to it. Its meaning is communicated by means of drawings, descriptions, myths, and so on, which are simply sign-vehicles themselves. The following two sections, discuss different approaches to the study of cultural-units that comprise the semantic component of semiotic systems.

3.7.1 Semantic Fields

It is already clear in Saussure's conception of language (cf. 3.5.5) that value of each linguistic term results from its position relative to all other terms in the language-system. Each term is defined by all the others which oppose and circumscribe it. Furthermore, Saussure maintains that the content plane or the signifieds (or ideas, as sometimes he calls) are organized in homologous fashion to the organization of sounds or the signifiers, i.e. language is structured isomorphically in two parallel planes.

Phonology studies the expression plane as a system (i.e. as an s-code, cf. 3.5.9) made of differences, and in accordance with Saussurean principle, describe its units (such as the phoneme, cf. 3.6.1.1) in terms of mutual differences (Malmberg, 1967, p. 13). By analogy to the phonological analysis of the expression system, structural semantics elaborates the general system for the semantic system (Lyons, 1977, p. 318). The value of a cultural unit (a sememe, cf. 3.6.1.4) is defined as it is placed in a system of other cultural units that circumscribe it, hence its value results from pure differences that the semantic s-code maintains (Eco, 1976, p. 73). The following example from Lyons (1968, p. 57) illustrates this:

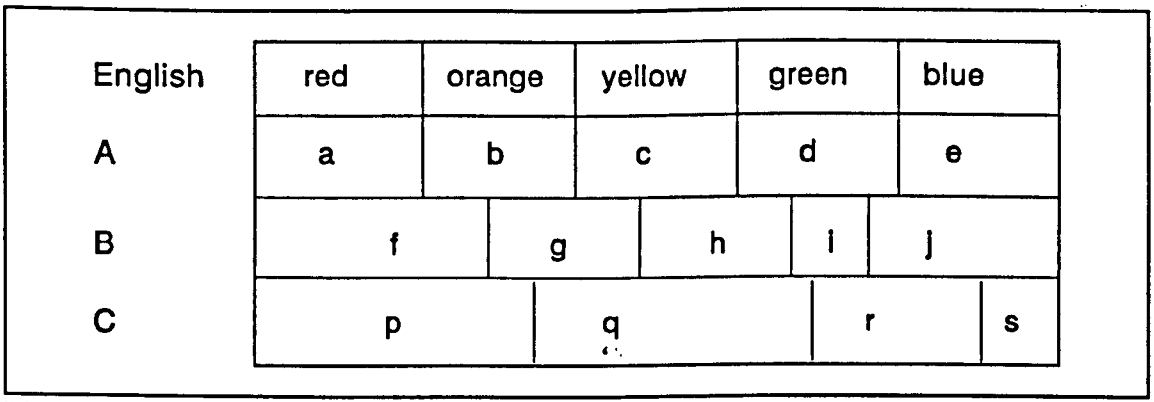


Figure 3.6: Value in Sign Systems (in Lyons, 1968, p. 57)

In this diagram English vocabulary divides the undifferentiated continuum of the colour spectrum into 5 colours. Hypothetical language A divides the spectrum in exactly the same way. Therefore, A is semantically isomorphic with English. B has the same number of divisions of the spectrum, however, the boundaries between the area of the spectrum covered by its terms does not coincide with English. C has just 4 terms for the same length of the spectrum, therefore differs from all the rest. This shows that the value (or meaning) of each colour-term is different in different semantic systems and result from the internal organization of the system that delimits the position of each term in terms of mutual differences. It is shown that various natural languages do in fact differ in their organization of the colour-terms (ibid, p. 56-59). A particular system (as an s-code; cf. 3.5.9), such as the colour-terms system that governs the organization of a group of semantically[21] related content-units or sememes is called a 'semantic field'[22,23] (Lyons, 1977, p. 268). The objective of the semantic-field approach is to define the semantic-space in its totality (i.e. as the form of the content-plane in Hjelmslev's sense, cf. 3.5.10) (Eco, 1976, p. 76).

3.7.2 Compenential analysis

An alternative approach to meaning in structural-semantics is that of 'componential analysis' (Lyons, 1977, p. 317) or 'compositional analysis' as it is sometimes called (e.g. Eco, 1976, p. 93)[24]. When the concept of sememe is introduced in 3.6.1.4, it was noted that Structural Semantics, especially those in the spirit of Hjelmslev's 'Glossematics' maintain that the semantic part of the language (i.e. the content plane) should be analyzed in a similar fashion to the analysis of the *expression plane*: as morphemes are made of phonemes, semantic units or sememes are made of smaller number of more general (universal) sense-components in direct analogy with phonemes (cf. 3.6.1.1). Thus, compenential-analysis sets out to define meaning or

[21] Whether *paradigmatically* or *syntagmatically*.

[22] Apart from the above example of the *semantic-field* constituted by the 'colour-terms', one can cite various others that are studied in detail, such as; kinship-words, zoological and botanical classifications, meteorological terms, and so on.

[23] A *sememe* can be member of a more than one *semantic field*.

[24] Although *compenetial analysis* of meaning was first developed independently of *semantic-fields* theory, it has many similarities with it and has been adopted in some of the later work in the semantic-fields theory (Lyons, 1977, p. 269).

sense of a linguistic-unit in terms of its constituting *universal* components (cf. 6.1.1). In America, a similar perspective is taken in analyzing meaning in structural terms, first by anthropologists and later by linguists. Katz and Fodor can be cited as the major exponents of this approach (Lyons, 1977, p. 318).

Although there are several important differences between various schools of compenential-analysis of meaning, some basic features common to all can be illustrated by the following example: the meaning or sense (i.e. sememe) of the lexeme "man" is composed of universal 'sense-components', such as; "MALE", "ADULT", "HUMAN", so on[25]. The sense components that make up a sememe (for instance, those correspond to the lexeme "man"), are variously called 'semantic components', 'semantic markers', 'semes', etc. (Lyons, 1977, p. 326; 1968, pp. 470-471) (cf. 5.7.1 and 5.7.2).

3.8 Metonymy/Metaphor

Consider the following imaginary sematic field (in Eco, 1979, pp. 78-79):

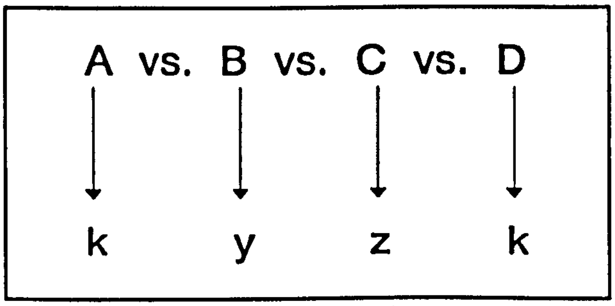


Figure 3.7: Metonymy/Metaphor

where; the horizontal line of the capital letters, constitute a paradigm of different sememes, and the vertical arrows constitute sememe to seme relations.

To name A by /k/[26,27] is a case of *metonymy*. For instance, A could be sememe <<king>> of the sign-vehicle /king/, and k could be one of its constituting semes that characterizes it, such as <<crown>>.

From the schematic representation of the semantic field in the above diagram, it can be observed

[25] How this formulation works will be described in more detail, when a specific approach to 'compenential-analysis' is examined, in connexion with design of IRS from a *semiotic* perspective in 5.7.

[26] Terms written between slashes hereby denote expression-units (sign-vehicles), and in double angular brackets denote content units (sememes) in accordance with the convention adopted in Eco (1976; 1979; 1984).

[27] According to the Peirceian model of *semiosis* (cf. 3.7) k which is a semantic marker (seme) of the sememe A, can, in turn, become a sememe on its own right and analyzed through other semes, of which A could be one of them (Eco, 1979, p. 78). See 5.7.3 for the discussion of IRS in terms of semantic markers.

that, k is also one of the semes of the sememe D. It is possible, by amalgamation through k, one can name A by /D/, this is a case of *metaphor.* A long neck being a property of both a beautiful woman and a swan, the <<woman>> can be metaphorically substituted for by the /swan/.

According to the above model, metaphor rely upon metonymy as contiguity in the code that connects the substituted terms. One can produce connections that has not been established before (culturally unknown), in this case it can be said that an 'aesthetic' message is communicated or an 'invention' has been made (ibid).

3.9 Language games and Speech acts

In 3.5.14, the extensional and intensional approaches to meaning is discussed. The extensional approach as developed by Frege, Russel, and early Wittgenstein, and others, maintains that meaning of a word consists of the object it refers to or denotes. For instance, the meaning of a term such as /hammer/ is the object it denotes (Bechtel, 1988, p. 19). Logical Positivists, who developed the extensional theory of meaning, claimed that ordinary language must be reformed because of its inherent deficiencies (i.e. ambiguities) and sought to develop a 'proper' language that could be clearly defined in 'logical terms' (ibid, p. 24).

The view that language can be understood by external objects, and ordinary language needs to be reformed, was criticised by Wittgenstein in his later work (1958) and by 'Speech acts' theorists.

According to Wittgenstein, the idea that words in a language can be defined in terms of its essential properties or the object(s) that they denote is fallacious. He demonstrates that there are many terms in language that do not have any 'defining' essential property, which differentiate them from other similar terms. This is according to him, not because of our inadequacy to develop formal systems that define language in terms of some essential properties, but because in language there are no essences. Wittgenstein, instead advocates that language must be understood by studying, how we *use it to do things with it* (cf. 5.5.5).

To illustrate, the variety of uses of language or variety of linguistics activities that takes place in ordinary language, Wittgenstein uses the term 'language games' (cf. 5.5.5) (ibid, p. 24). By comparing language use to playing games, Wittgenstein asserts that there are *undefined* number of games that may be played, some of which have fixed rules, some of are alterable, some of them are invented as we go along. Language can be, for example, used to make; 'assertions', 'prescriptions', or to; 'instruct', 'order', 'interrogate', so on[28]. When using language, participants in the game, draw the boundaries of the game, choose the appropriate rules for the situation which can be changed in due course as required (Brown, 1974, p. 34).

Wittgenstein's view on ordinary language is shared, by speech acts theorists; Austin (1962), Searle (1979), Grice (1989), who analyze language as a kind of 'action'. Speech act theorists maintain that, instead of striving to reform ordinary language, what is needed is to focus

[28] For further discussion of 'language games', and its application in IR context see 5.5.5.

carefully on how language 'functions' in practice (Bechtel, 1988, p. 27)[29].

3.10 Subject of Semiotics

As we have mentioned in 3.1, semiotics deals with everything that can be taken as a sign by somebody, i.e. as Morris puts it: "something is a sign only because it is interpreted as a sign of something by some interpreter....Semiotics, then, is not concerned with the study of a particular kind of objects but with ordinary objects insofar (and only insofar) as they participate in semiosis" (in Eco, 1976, p. 16).

Therefore, semiotics deals with all sorts messages emitted by variety of organic sources (humans, animals and plants) or by some of their component parts (such as information/message transmitted in cardiovascular system) and inorganic (artificial, such as machines, or natural, such as smoke signalling possible fire) objects (Sebeok, 1976, pp. 2-3).

It is possible to conveniently divide the semiotic sphere into two domains. Messages are interchanged either in *human systems* or in *other living systems.*

The human semiotic systems are studied by *Anthroposemiotics*, and includes human computer communications.

Zoosemiotics studies the animal communication systems. Anthroposemiotics and zoosemiotics compromise the two main divisions of semiotics with common certain essential features, differing especially in the fundamental role played by 'language' in the former.

One can also mention a third domain, namely, *endosemiotics* which studies cybernetic systems within the body. The genetic code plays a role in this field, comparable to that of the verbal code in anthroposemiotics, although it is generally accepted that coding and transmission of information inside the body is very different than that of outside the body (ibid).

One can further divide the semiotic sphere according to *terrestrial* and *extraterrestrial* communication. The latter is explored by mathematics, exobiology, radioastronomy and such (ibid).

It is also possible to classify signs according to being intentional or non-intentional. Intentional signs are produced in human communicational acts. Non-intentional signs could be one of two types: *a)* physical events such as; smoke indicating a possible fire, a trace left on a track, indicating a passage of an animal, or red spots on a patient's face indicating, e.g. 'measles' *b)* human behaviour not intentionally emitted by its sender, such as; gestural behaviour, signalling the cultural origin of the gesturer (Eco, 1976, pp. 15-18).

[29] See Winograd & Flores (1986, pp. 17-20, 112-114) for a good argument against analytic (extensional/positivist) definitions of meaning from the point of view of *speech acts* theory, especially as it is developed later by Habermas.

Chapter 4
SYNTACTICS OF IRS

To benefit from semiotics in analyzing the documentary information retrieval situation in the context of information science it is necessary to treat IRS as sign systems and IR as a communication process. This entails the identification of the basic structure of semiotic systems in archetypal IRS[1] taken as a 'communication medium'[2].

It appears as if there is a general consensus on the idea that Information Science in general and Information Retrieval in particular is about transmission of information, or 'human communication' (e.g. Ruben, 1992; Vakkari & Cronin, 1992; Belkin, 1980; Belkin & Robertson, 1976).

It is usually assumed, explicitly or implicitly, in many IR work that, the objective of IR is to satisfy the user's information need by presenting the user those document(s) that match best to her or his query[3]. Robertson & Belkin (1978a, p. 93) puts this succinctly: " (a) A user, recognizing an information need, presents to an IR system ... a request, based upon that need ... (b) The task of the IR system is to present the user with the text (or texts) which it judges to be most likely to satisfy the user's information need, based upon the request put to the system. (c) The user examines the text ... presented by the system, and her/his need is satisfied completely or partially or not at all. The user's judgement as to the contribution of each text in satisfying the need establishes the usefulness or relevance of the text to the need".

The above description implicitly assumes the following model of communication: the user foremost, expresses the whatever information need that s/he may have in natural language which is then expressed in the query language of the system if this is needed; this description refers to some objects (i.e. the documents) in the database; and these objects in turn may or may not resolve the information need of the user. The assumption being that, the query terms are the description/expression of some information need which is (roughly) equivalent to the documents that the user judges relevant for her/his need.

If IR is a communication process as it is usually assumed in the IR literature, then the above characterization of the IR process should hold true in terms of semiotic categories. As semiotic analysis of the IR situation performed in the rest of this chapter and the subsequent chapter shows, this is in fact not the case. It is the argument of this chapter (and chapter 5) that the above description of the IR situation cannot be upheld in the light of a semiotic analysis.

The most important feature of the above depiction of the IR process is that, it is viewed as the

[1] Such as described in IR text books, e.g. Ashford & Willet (1988), Lancaster (1979), Salton & McGill (1983), etc.

[2] That is, a 'sign system' used for *human* communication purposes. Similar to the Andersen's (1986, pp. 68-70; 1990) conception of 'computer as a communication medium'. Winograd & Flores (1986, p. 123) proposes a similar view to this.

[3] Not necessarily the 'query statement'. See section 5.5.1.

problem of finding that of the equivalent (or most close approximation) of the 'information need' in the document collection. Furthermore it is assumed that, the information need (therefore the relevant documents) is foremost expressed or described in natural language or at least translatable to a verbal language (this may need to be further translated manually or automatically into the query language of the system).

Thus Belkin writes, referring to the 'ASK' hypothesis : "One of the assumptions underlying this project is that a state of knowledge can be represented by a network of associations between *words, standing for concepts*" [my *emphasis*] (Belkin et al, 1982a, p. 147).

This description[4] closely approximates, Richard and Ogden's sign structure (cf. 3.5.1) :

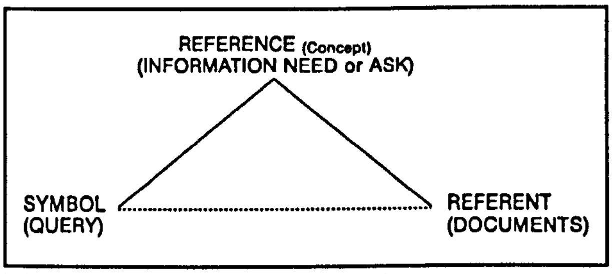


Figure 4.1: The Q-I-D Model

I would like to pick up the following points from the above description which assumes (implicitly) that:

a) the code underlying the generation of a query or request submitted to an IRS, and the code underlying the generation of natural language discourse, are one and same.

b) the query describes the relevant documents.

c) the relevant documents satisfy the information need which gave rise to the query in the first place.

We need to examine the above three points in some detail to see whether they correspond correctly to the semiotics of documentary information retrieval situation.

To start off with point (a), if we accept this assumption, then we have to accept that in a request such as "I would like to retrieve some articles on *artificial intelligence*", the term 'artificial intelligence' has the same meaning as the one which occurs in a natural language discourse such as; "Any student of modern thought, from literature, art history, zoology, theology, *artificial intelligence*, should read it straight through, and then pick at it here and there ...".

Anybody who has ever used an IRS would agree that this is not so. The meaning of the term 'artificial intelligence' in the query is hardly the name of the academic field of specialization called 'artificial intelligence' *alone*, it is rather one, or several, or all of the things that is related to the field of specialization called 'artificial intelligence', (methods, systems, tools, applications, theory, etc), including the name of the field of study itself. In the second case (i.e. natural language discourse) it is obvious from the context that it is the 'name' of the intellectual field that one is referring to, which is most certainly not equivalent to the meaning in the first situation.

[4] It will be referred as the 'Q-I-D' model (Query-Information need-Documents).

Therefore, it is necessary to answer the following question: "What is the *value* (meaning) of a term in an IR 'situation'?". It is the objective of this chapter to answer this question.

The remaining part of this chapter discuses the *syntactics* of IRS, i.e. the formal structure of the system as a 'system of signification' and how various constituting parts relate to one another[5]. This is the subject of 'theory of codes' or 'semiotics of signification'[6] as would be recalled from section 3.5.9.

In chapter 5 the actual process of production of signs to communicate is analyzed. It is therefore, not only interested with the structure underlying the communication process but also with the act of using signs. This, as would be recalled from 3.5.9 is the subject of 'theory of sign production' or 'semiotics of communication'.

4.1 Expression/Content, Form/substance in IRS

The objective of chapter 4 is to analyze IRS as a semiotic system and determine its structure and constituting parts.

To do this, it is necessary to identify its various components and the functional relations between them which altogether constitute 'IRS as a code' and enable it to function as a 'sign system'.

The following need to be determined in order to explicate the structure of a semiotic system:

- the expression and content planes
- the structure (s-code) of both planes
- the relation between the two planes (the code)

It is the objective of sections 4.1.1 and 4.1.2 to explicate the expression and content planes in IR. Sections 4.2 and 4.3 deals with the structure of expression and content planes taken as s-codes (cf. 3.5.9). The relation between the two planes which enable IRS to function as a sign system or code is analyzed in section 4.5.

4.1.1 Expression and Content in IRS

It is the purpose of this section to explicate the content and expression planes (cf. 3.5.10) in a typical IRS formally, deducing the structure of the correlative system from more primitive axioms and rules that has been laid down by the semiotic methodology that has been developed so far.

It has been noted in 3.5.10 that the formal model developed by Hjelmslev treats both planes to

[5] It is hoped that the point "a)" posed above will be answered by the end of this chapter. Points "b)" and "c)" will be left to chapter 5 for thorough treatment.

[6] It is however not particularly useful to dissociate 'semiotics of signification' from 'semiotics of communication', i.e. semiotics that studies the actual processes of sign production, completely, and therefore the analytic object of chapter 4 (i.e. system structure) overlaps partially with that of chapter 5 (production of signs).

be of homologous in structure; each plane having the structure of form/substance/purport, i.e., a given purport moulded into different sign substance by the sign forms, as established by the code (Posner, 1992, p. 42). The sign forms are the invariants of a given code, and are determined by what is called the commutation test, as already noted in the same section.

To establish the invariants of the expression and the content planes of an IRS, it is necessary to apply the commutation test (cf. 3.6.1). To apply the commutation test however, we need to know what constitutes, at least, one of the planes so that we can apply the test to its elements and determine what corresponds to them on the other plane.

It is fairly obvious that in IR, query, or more precisely the terms that constitute a query, is the expression of some content. Therefore, we can assume that expression plane in IR is constituted by the query terms. This can be treated safely, as a postulate of the current theory (the semiotic model) rather than a theorem or hypothesis.

Having postulated this, we can now apply the commutation test to the elements of the expression plane. The question to be asked whenever applying the commutation test is; whether a change occurs when a portion of the text is pulled out and substituted with another, and if there is a change what constitutes it. In IR situation this can be translated as; if a query term is replaced by another, is there a corresponding change in the system's 'state' and if there is, what this change entails in terms of the structure of the system.

Substitution of a query term with another would result in a different set[7] of retrieved documents. The contrary would mean either that, the terms substituted for each other are not invariants[8] or they do not belong to the system at all (not indexed). From this analysis, it can be concluded that query terms are correlated with the retrieved documents, therefore the retrieved documents function as the content of the query terms. This will be discussed in detail in 4.2.3.

It is worth noting at this stage that the commutation test shows content of a query is the retrieved documents and not the relevant documents[9].

Having established the content plane in IRS, it is now possible to apply the commutation test on the document collection and observe its effect on the expression plane. Substituting a retrieved document with a not-retrieved one will always result in a corresponding change in the expression plane (i.e. query terms), validating our previous conclusion.

[7] The argument, without fundamentally changing its line of reasoning can be extended to 'ranked' systems, where the output, technically speaking, is not a set but a ranked list. See 4.1.2 for further discussion of this point.

[8] That is, absolute synonyms, in the sense that substitution of one for another results in exactly the same set. This is the case for instance, when different inflextional forms of a 'lexeme' is represented by one of its forms or (more likely) by the stem of the lexeme; e.g word forms 'office' and 'offices' can be represented by the stem 'offic'. The commuted terms in the example could also be 'stop words' which would result again in no change in the retrieved set.

[9] As van Rijsbergen (1989; p. 78) says: "... the computed relationship between a document and query is not one of relevance". Speaking *like* Deleuze and Guattari (1987), 'relevance' is *not* the *object* of the *retrieval machine* (cf. 6.4.2), or speaking *like* Derrida (1974), it is its *supplement*.

It is worth noting once more that, there is no correlation between the relevant documents and the query. It can be observed that, substituting a relevant document with a non-relevant would not result in a corresponding change in the expression plane. If the division in the content plane (i.e. the invariant units of the content plane) were between relevant and non-relevant documents, the commutation test would result in a corresponding substitution in the query terms, which is clearly not the case.

4.1.2 Form and Substance in IRS

As noted in the preceding paragraphs, Hjelmslev's model of the sign-function, which is the most formally developed of all the followers of the Saussurrean lead, divides each plane of the signification process into three successive layers of purport/form/substance.

The preceding section identified the documents as the content plane and the query statement as the expression plane in IR. To arrive a more precise model of the sign-function contracted by IRS, it is needed to analyze each planes in terms of purport, form, and substance.

Purport on the expression-plane is the matter that a sign-vehicle[10] is made of at the moment of its material instantiation in an act of sign-production. In speech for instance, purport is the sound waves that a human speaker produces at the instance of uttering a word. In written text, it is the graphic material that are left as traces on paper. On the content plane it is the stuff that thought is made of in the conceptualist or mentalist version of the model. As it is indicated in 3.7 however, the semiotic approach adopted here treats the content units as cultural units that are *materially, physically* within our grasp. Content-units, like expression-units, are material sign-vehicles that correlate with other sign-vehicles. Therefore, they are similarly made of sound-waves, graphic material, etc.

In IR we have seen that documents are correlated with query-terms. Documents function as the content of the expression which is the query statement. The material that both the expression and content are made of in IR, is some graphic material. More accurately, documents and query terms are represented in the system, as binary digits or bits, and at the interface or the output device, such as a monitor, as pixels or luminous material.

Form is the function that transforms the purport into concrete, material occurrences, that is the sign substance. On the expression plane, form determines invariant features of the expression-substance that is necessary in order to distinguish it from all other expression-substances that can be generated within a given code. On the content plane, similarly, form determines the pertinent, invariant features that a 'content-substance' must possess . In the framework of semiotic perspective adopted in this dissertation, content-form imposes structure to the expression-substance that *function as* the content of some other expression-substance, i.e. it determines the invariant features that a content-substance must possess in order to be identified as a distinct[11] unit.

In IR situation, expression-substance are the query terms or more accurately the search statement. Search statement is the actual physical occurrence of an expression-vehicle at the

[10] That is expression-substance, see the discussion of form below.

[11] Or, as a *token* of a *type* (see, 4.2.2.1 for definitions of *type* and *token*).

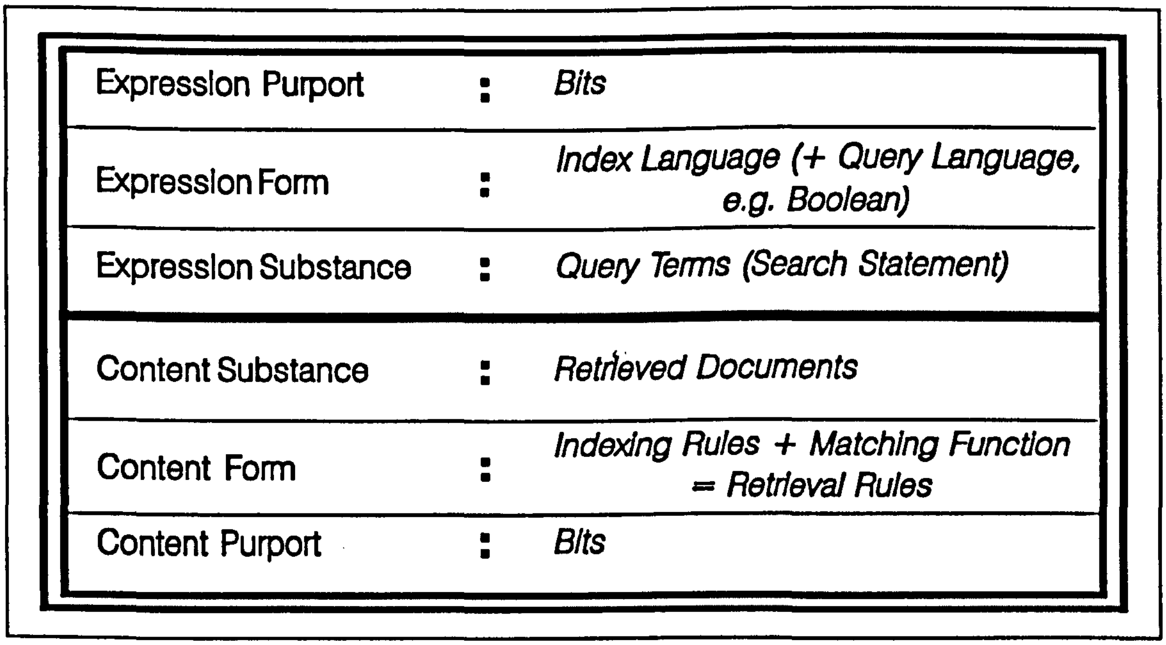
moment of uttering or expressing some content. However, only some features of the expression-substance or the search-statement is pertinent for the functioning of the IRS. This is because, almost all IRS involves some sort of vocabulary control, where only some terms, more accurately, stems of some terms are indexed, therefore, available for searching. Since only searchable parts of words and word forms are pertinent for the retrieval purpose, they constitute the expression-form in IR. It can be said therefore that, index language comprise the form of the expression plane in IR. One can use the tool of commutation test to verify this as discussed in 4.1.1. It should also be noted that apart from the terms in a search statement, logical operators that determine the proximity relations between the terms constitute the expression-form. In some systems difference between, for instance, 'adj' and 'and' may not be pertinent, whereas in some others, it may.

As it will be shown in 4.2.3, the form of the expression and content planes are isomorphic in IRS. For the time being though, we can analyze the form of the two planes separately.

As it is shown in 4.1.1, the pertinent distinction on the content plane in IR is between, *retrieval/not-retrieval*, in other words between *retrieved* and *not-retrieved documents*. It is not for instance between those documents that are marked by the user as relevant versus those marked as not-relevant. From this argument, it can be concluded that the form of the content plane in IR is characterized by the retrieved v. not-retrieved distinction. The content-substance in IR is the documents retrieved by particular query terms. The content-form determines the pertinent features of the content substance and divides the document collection into retrieved and not-retrieved sets[12].

We have mentioned above that the forms of the expression and content planes in IR are isomorphic. This is a very important property whose pivotal role in functioning of IRS will become clear in the rest of this chapter, and indeed in chapter 5. It is however useful to spend some thought on it at this stage. It is concluded above that, the content form in IRS is retrieved/not-retrieved distinction. It can be shown very easily however, this distinction depends in the final analysis on the terms that index the individual documents, i.e. on the index language or more accurately the indexing rules and the matching function (i.e. inference rules or retrieval algorithm) which selects the documents from a collection. The indexing and retrieval rules determine therefore the pertinent features of a content-substance, which is/are the document or documents that are indexed by a particular index term. This is the very same form which impose structure to the expression-substance as it would be recalled from the preceding paragraphs. Figure 4.2 summarizes these findings.

[12] The *ranked* output systems, although does not necessarily divide the document collection into mutually exclusive retrieved/not-retrieved sets, rather, order the whole collection in a some specific way, what matters in practice is the availability of a document rather than another for the user. The whole purpose of a ranked list is to make available a document to a user in preference to some other, therefore, it can argued that in this case the *pertinence* is between *availability/not-availability* or *presence v. absence*, which is hardly different than the *retrieved/not-retrieved* dichotomy for most *practical purposes*. The contrary argument would defeat the *raison d'être* of the ranked output.



Expression Purport	:	*Bits*
Expression Form	:	*Index Language (+ Query Language, e.g. Boolean)*
Expression Substance	:	*Query Terms (Search Statement)*
Content Substance	:	*Retrieved Documents*
Content Form	:	*Indexing Rules + Matching Function = Retrieval Rules*
Content Purport	:	*Bits*

Figure 4.2: Purport/Form/Substance in IRS -- System Structure

4.2 The Value of Sign in IR

To explicate the relation between the two planes identified in the previous section, it is necessary first to analyze the expression and content planes in more detail. To do this we need to have a closer look at the classification of signs. In 3.5.3 a classification of signs as proposed by Peirce was given, the main divisions being: Symbol, Icon, and Index.

When intension and extension are discussed in sec. 3.5.14, it was pointed out that in the context of the theory of codes, that is, how sign-vehicles are correlated with content units, there is no need to recourse to the referent or the object that is 'out there'. In fact it was noted that if a general theory of signification is to be developed, one should avoid the referent altogether, even it is for the simple fact that many signs do not have any referent.

The idea of referent is certainly useful in certain circumstances, for instance, when one refers to or mentions a particular object. This is covered by theory of mentions, which deals with signs that are used to name things, describe actual state of the world, to assert that there is something and it is so and so, etc. (Eco, 1976, pp. 161-162). The theory of codes or signification does not involve with the actual state of the world as such . As noted in 3.5.14, signification is not concerned with the truth-value of sentences or propositions, which is the subject of logical (extensional) semantics[13].

In Peirce's trichotomy however, both categories of Icon and Index involves explicitly the idea of an external referent. This is criticised heavily by Eco (1976; 1979; 1984) and the following section is intended as a brief discussion of this criticism.

[13] A similar argument can be found in detail in Winograd & Flores (1986, pp. 17-20, 27-37, 54-79).

4.2.1 The untenable trichotomy

The title of this section refers directly to the title of Eco's section (1976, p. 178) on criticism of Peirce's trichotomy on the grounds that the idea of referent is present in their definitions as a discriminating parameter.

Eco's main points of objection to extensional definition of signs have already been discussed in section 3.5.14. Suffice to say here that whenever one tries to point an object precisely, even at the act of mentioning, one encounters with yet another sign(s) and another level of semiosis which precedes the act of 'perception' of the object (ibid, p. 165). All this become clearer when we discuss Eco's ideas in detail in the sections that follow.

It is worth to note here however that, the idea of a text or a document as an object is even less tenable in the light of contemporary semiotics and philosophy of language as discussed in detail in section 4.2.3. This alone, forces the course of methodological vocation of this dissertation to follow the theoretical lead of Eco as developed in his volume 'A Theory Of Semiotics' in discussion of IRS as semiotic systems.

4.2.2 An alternative classification of Signs

Eco in the same volume proposes an alternative classification for signs which avoids the shortcomings of the extensional definition. It is based on the concepts of 'type' and 'token' which originates from the work of C.S. Peirce.

4.2.2.1 Type and Token

A Sign-type is the abstract class of all Sign-tokens which (by some criterion[14]) belong to the same type (Bar-Hillel, 1964, p.41).

A Sign-token is the concrete (physical) instantiation on some specific occasion of a particular Sign-type (Cherry, 1968, p. 308).

The following example from Lyons (1977, p. 13) illustrates the concepts of type and token: on every occasion on which the word 'reference' occurs, the letter 'e' is instantiated four times, that is, on each occurrence of the word 'reference', there are four 'tokens' of the 'letter-type' 'e'.

Similarly larger syntagmatic chains such as words-tokens are reproduced from general types. In the two sentences 'Whatever happens Berlin will remain Berlin' and 'Berlin is situated in Germany' contain the same word-type 'Berlin', appearing in three different tokens (Posner, 1992, p. 42). In natural language systems, as in some other systems, expression functives are reproduced according their type. They are reproducible indefinitely according to their model (cf. 5.3.3).

[14] As determined by the code.

4.2.2.2 Type/token ratio

Type dictates the essential properties that the token must posses in order to be accepted as a valid reproduction of the type, thus ensures its recognition. Tokens of the same type can possess individual characteristics (free variations) as long as pertinent ones are retained. The commutation test mentioned in 3.6.1 is used to establish the pertinent (invariant) features. We can distinguish between two kinds of replicas; *absolute* and *partial* (Eco, 1976, p.180).

An absolute replica pertains all the properties of its type, thus, also called as double. Two objects produced exactly the same are doubles of each other, such as two cars of same make and model. Traffic signs are an example of signs which must duplicate their type without any variation, thus, they are doubles or absolute duplicative replica too.

Partial replicas need only reproduce some of the features of their type, those established by the code as invariant. Signs in natural languages are this kind. An uttered word is not a perfect double of another word of the same type. Variations are allowed as long as it possess the invariant features as dictated by the type. Phonemes, words, ready-made expressions, etc. are produced according to the pertinent features of their type, therefore, they are all partial replicas (ibid, p. 182).

Every sign is therefore, produced according to some type/token ratio, which ranges from perfect reproduction of the type to loose reproductions which allow for a great deal of free variations (such as, images on playing cards) (ibid, pp. 178-183).

4.2.2.3 Ratio facilis/difficilis

Eco's (1976) classification of signs rests on the distinction between two particular kind of type/token ratio, referred to as 'ratio facilis' and 'ratio difficilis'.

Ratio facilis covers the cases discussed so far, where an expression-token is generated according an established expression-type. All signs replicable according their type are governed by ratio facilis, phonemes, lexemes, etc., are the prime examples: "There is a case of *ratio facilis* when an expression-token is accorded to an expression-type, duly recorded by an expression-system and, as such, foreseen by a given code" (Eco, 1976, p. 183). In ratio facilis the expression is correlated to its content arbitrarily[15].

Ratio difficilis governs the production of signs where the generation of expression is affected somehow by its content. "There is a case of ratio difficilis when an expression-token is directly accorded to its content, whether because the corresponding expression-type does not exist as yet or because the expression-type is identical with the content-type. In other words, ... *the expression-type coincides with the sememe* conveyed by the expression-token" (Eco, 1976, p. 183).

In ratio difficilis the expression is not arbitrarily correlated with its content. The correlation is so to speak motivated. Motivation in this context means that some of the 'markers' (cf. 3.7.2)

[15] That is by *convention*, in the sense that there is no *motivation* or *casual* link (cf. 3.2). Compare with ratio difficilis below.

of the expression-token is same that of the corresponding sememe[16].

The case of ratio difficilis as proposed by Eco is to replace Peirce's categories of Icon and Index. It replaces in the definition of Icon (and Index) the motivated resemblance of the sign to the object (referent) by the homology (isomorphism) of the expression-type and the content-type (the sememe), thus eliminating the necessity of actual presence of the referent in the definitions.

Ratio difficilis governs so called iconic signs, where there is said to be motivated relation between the sign and its referent. An example of a sign governed by ratio difficilis is the imprint of the foot of an animal on ground. Even in this case of apparent motivation between the imprint and its referent, the foot of the animal, i.e. the relation of cause-effect between the sign and the referent is culturally pre-established by series of mentions, therefore, meditated through other signs that function as their interpretant[17] (ibid , pp. 222-223).

4.2.3 Type/Token Ratio in IRS

To establish the value of sign in IR, it is necessary to determine its type/token ratio.

Very few researcher in IR would probably object to the idea that the terms that constitute a query or a request are the expression units that are used to communicate some meaning or content in the retrieval process (cf. 4.1). The next task is to identify the system or more accurately the s-code[18] underlying the generation of the expression units.

We will start with a simple observation: the query is produced (formulated) according to the form[19] of the documents that it is to be retrieved. That is, the pertinent features from the documents that are *thought to be in the database* are selected to produce the query/request[20]. The expression in IR is, thus, accorded to the (physical) form of the documents (thus, this is a 'topological' condition[21]) in the database[22]. In other words, the documents set the model

[16] This is the case usually, when the markers of the sememe are describing spatial relations. This is called topo-sensitivity (Eco, 1976, pp. 184-187).

[17] Interpretant as defined in 3.5.2.

[18] It would be recalled from section 3.5.9 that, s-code in the semiotics terminology we have adopted differs from code in that, it is concerned with the syntactics of either of the expression or content planes, whereas, the code governs the correlation between the two planes.

[19] *Form* denotes the spatial properties of the documents, i.e., 'structural' organization of the text into paragraphs, such as, title, author, abstract, subject, keywords, etc., i.e., division of the scientific discourse into logical discrete units. It also denotes the (form of) individual signs (words) that make up a text (cf. 4.1.2). The syntagmatic relation between the terms also determines the *form* of a document.

[20] This, no matter how trivial it might seem to be, is *the* single most important characteristic of documentary information retrieval situation that needs to be explicated further in order to advance our understanding of functioning of signs in IRS.

[21] That is, the spatial disposition of the terms in documents determine the form (spatial disposition) of the query statement.

(content-type[23]) from which the expression-tokens are replicated/produced[24] (therefore, this transformation is governed by ratio difficilis).

But is it really necessary to have the documents that the query is supposed to refer in the database for this system to function as described above? The answer is clearly 'No'. Even though, the required documents may have not been indexed in the database, the query is generated from the pertinent features of the supposed documents.

One can thus conclude from the above reasoning that, the query in IR is generated according to a content model that refers to the objects in the collection, which are the documents (such is the position of the 'Q-I-D' model mentioned at the beginning of the chapter). This is a rather overhastily reached conclusion however, as one has to answer satisfactorily, what is the exact difference between the content model and the documents that are supposed to be represented by it.

To accept the model depicted in the preceding paragraph is to assume that there exists a content model out there, and the problem is to find the true value of it among the documents to which it refers to, which again have an autonomous existence. This, as it would be recalled from section 3.5.14, is a position which assumes that the referent is a discriminating parameter in determining the meaning of signs (i.e., extensional semantics), which has been criticised as unsuitable for the semiotic phenomena i.e. the process of signification, from the perspective of the theory of codes (cf. 4.2.1).

Furthermore, to assert that documents are indeed objects, is to assert that they can be described or named unequivocally. This is of course against all the wisdom generated by both linguistics/philosophy of language and Information Science itself. As a matter of fact Information Science teaches us that there is no single way of describing documents, otherwise indexing them should also be straightforward. A document which is a text is not a closed, self sufficient entity, but rather a complex discourse which opens up at various levels to other texts in an intertwined web of citations and references. This is what is known as 'intertextuality' in contemporary literary criticism and philosophy (see for instance, Kristeva in Moi, 1986, pp. 75-77)[25].

[22] The structure of the database itself impose additional structure to the logical organization of the scientific manuscript. The original manuscript is usually transformed by abstracting, editing; controlling the vocabulary, and imposing indexing rules, etc. This determines the actual form of the document, and the query is accorded to this form ultimately. However, this can be thought as a special case of full-text indexing of the manuscript as it is originally published. Our analysis is not affected by either of the cases, therefore they will not be differentiated in the rest of the dissertation, unless the difference becomes pertinent for the analysis.

[23] As determined in section 4.1.1 that documents constitute the content plane in IR.

[24] See section 5.6.1.1 for the discussion of *replication/invention* dialectic in production of expression in IR.

[25] See Froehlich (1989b) for a similar argument against the view that documents are self-contained entities from the perspective of the *phenomenology* of Husserl and Heidegger (especially p. 63). See also Brier (1992, especially pp. 102-103) for criticism of the idea of objective information sitting "out there", from a similar point of view. One should also note Derrida's (1974) contribution in dismissing the idea of *self-sufficiency* of texts (the *metaphysics of presence*) (cf. 4.3.2.1).

Text in the above sense is not only the verbal text as it is commonly understood, but whole sort of objects, artifacts, systems, the experience of using them, etc. When we speak of texts referring to other texts in intertextuality, it should be understood as verbal texts referring to other verbal texts (written and spoken) and to non-verbal texts, and vice versa. For instance, a written text refers to other written texts, but might also to spoken ones such as a lecture as well as to non-verbal ones such as, a system or practical experience of using a system, and doing something with it; just as a lecture refers to written texts, artifacts, systems, and so on.

What constitutes the database is therefore, all sort of texts, producing a complex discourse. This interwoven complexity constitutes practically a dense continuum from which pertinent units are picked-up[26] to compose the expression, i.e. the query statement (cf. 4.3.2).

The content model and the documents in the database are therefore, one and the same thing[27], they are interpretants of the query terms, which themselves are simply signs. This is exactly the idea of (unlimited) semiosis[28] (cf. 3.5.2) as conceived by Peirce.

The query terms are generated according to the form (therefore, they are topo-sensitive) of the pertinent features of the documents, which refer to the documents in their absence. Once the query is submitted, it results in the retrieval of some documents (or none), and now some text is retrieved, it becomes a sign-vehicle (expression) which refers to something else which is absent. A more formal and thorough description of this will be presented in 5.6 .

One has to accept that, any effort to establish an analogy between a physical object and a document is rather a pointless exercise that does not lead to any useful definition of text. Whenever one tries to recover something more substantial than that of the text itself, one come across another set of signs which act as its interpretant, and this process continues to *ad infinitum*. One has to conclude that, there is nothing more (substantial) beyond *unlimited* semiosis.

From the above two points; *i)* the query is produced according the form of the (supposed) documents, *ii)* the documents are the interpretants[29] of the query terms; the semiotic model of IR starts to take shape.

To summarize this section; the type/token ratio in IRS is found to be that of ratio difficilis, in which, expression-tokens (that is the query terms) are produced according to the content-type[30] that is set by the document (text) collection, which in turn, as a whole constitutes a discourse on a given subject. At this point, we can go back to the remark made at the beginning of the chapter. The content of an expression in an IRS is not an atomic unit (sememe, cf. 3.6.1.4 and

[26] The selection of the pertinent units is governed by certain transformational rules, which are prescribed culturally (cf. 4.3.1.2).

[27] In the sense that the documents are inseparable part of a dense continuum (cf. 4.3.2.1) which altogether constitute the content-model in IR.

[28] Metonymic substitution (cf. 3.8) of a sign for another, *ad infinitum*, as interpretants of each other.

[29] See the discussion of IRS as encyclopedia in 5.7.3.

[30] In IR expression and content types coincide.

3.7.2) with well defined boundaries[31], but a text corresponding to a discourse. Thereforë, the expression units of IRS are incommensurable with the units of the natural language discourse, namely, words, lexemes, morphemes, and so on. (see 5.4.1 for a further discussion of this). In section 4.5.3 a more detailed comparison of the two modes is presented.

4.3 The Content Model (semantic model)

In 4.2.3, it is determined that, the mode of sign production in IR is that of ratio difficilis, and a general discussion of the process which leads. to production of signs (more accurately, expression units) is presented. The present section deals with the structure that underlies this process. First, in section 4.3.1 a general model developed by Eco is examined, and in 4.3.2 this is appropriated for the discussion of the processes which lead to production of expression in IR situation.

4.3.1 The transformative process

The general model of the transformations underlying the production of signs is given in Eco (1976, pp. 245-261).

According to this model (fig. 4.3), expression tokens are accorded to a semantic model, either arbitrarily as in the case of ratio facilis, or by means of conventional rules of similitude, as in the case of ratio difficilis. In the following sections, transformations leading to creation of the semantic model is discussed.

[31] In the sense that, sememes are made of semes (3.7.2, see also the simple model of a semantic field in 3.8, in this connexion). In 5.7.2, when *encyclopedic model* of IRS is discussed, a more detailed definition of the semantic system will be given.

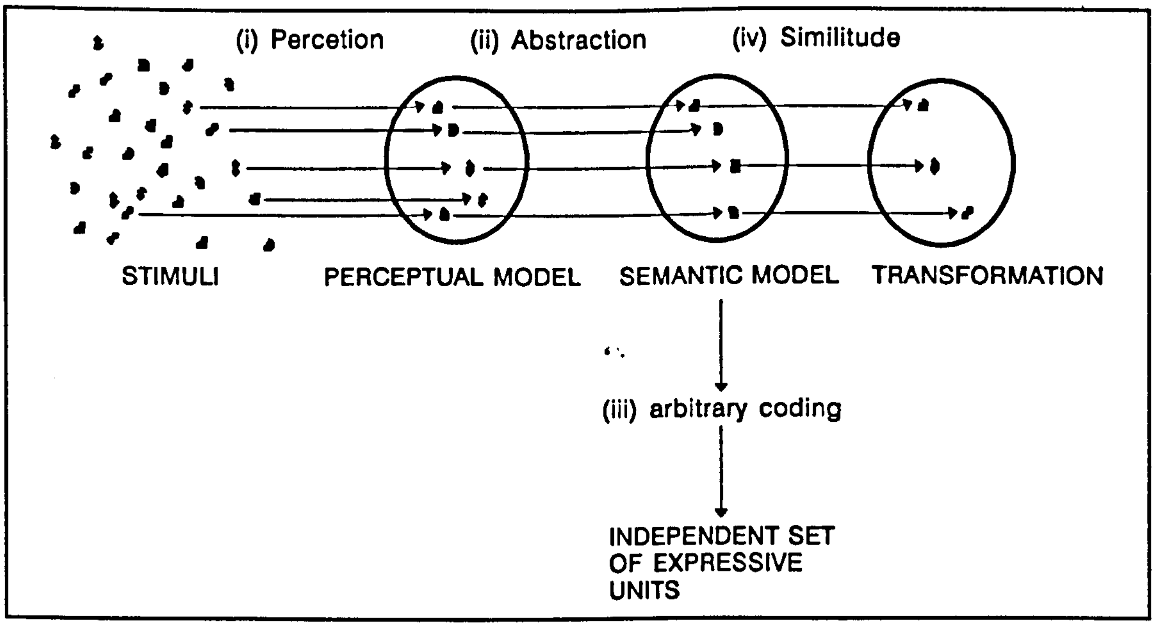


Figure 4.3: Transformations in Sign Production (in Eco, 1976, p. 251)

4.3.1.1 Mapping from stimuli (from stimuli to the model)

This section is intended as a summary of Eco's general model of sign production. Eco's model is based on three layers of transformations, each layer articulating the previous one. At the bottom of the hierarchy there is disorderly state of perceptive stimuli, constituting an unorganised and continuous field of perception.

This is the case for example, when one receives visual impulse from an object, say the sun. The sun itself is visible to us through the rays of light it emanates. The sun rays constitute a continuum of electromagnetic field. Our experience of sun light however, at least for most people, based on looking at the sun through partially closed eyes, which results in the perception of sun as a shining spot in the sky from which bright rays of light emanate.

This is the 'perceptual model' of the sun resulting from our perception of the object by our senses. This is not to say that even our perception model is not conditioned by cultural conventions. Had there been a culture which did not consider intensity as an important property of light, their perceptual model of the sun would not base on the brightness of rays but some other property which would mark it as different from other celestial bodies.

The iconic representation of the sun as a circle with short lines emanated from its centre however, is not directly based on this perceptive model. To represent the sun as such, this perception needs to be transformed by abstraction into semantic (logical) model of the sun as a spherical body radiating light waves which are thought to travel in straight lines. This (semantic) model is a result of our understanding of celestial bodies in the cosmos, based on our scientific knowledge. The scientific knowledge enables us to assign the sun the semantic properties of being spherical and light rays travelling in straight lines. This is a highly abstract model however, as, whether light is taken as quanta or waves, is certainly not made up of straight lines.

Once the semantic model is established it is a matter of graphic conventions to represent a sphere as a hollow circle and radiating light rays as co-centric short straight lines.

In this sense iconic representations are related to their referent (object), however this relation is mediated through cultural conventions, i.e. other signs. It is therefore, correct to say that iconic representation of the sun is a schematic representation of a semantic model, which itself is a schematic representation of the actual celestial body that we name as the sun. The image of the sun is motivated by the abstract representation (the semantic model) of the sun, which itself is a result of cultural conventions. Since it is a result of cultural conventions, the similitude between the sign-vehicle and its referent need to be learned to be recognized.

It is possible to give plenty of example from history of art to illustrate how cultural habits and conventions condition the representation of various objects and living creatures in different epochs and cultures (see e.g. discussion in Eco, 1976, pp. 204-205).

4.3.1.2 From the model to the expression

The so called iconic code establishes equivalence between a pertinent unit of the semantic model (system), and a pertinent unit of a graphic system.

The difference between this sort of coding, and a code such as, the natural language is that, in the case of ratio difficilis some of the markers of the expression units are the same as that of the semantic model, whereas, in the case of ratio facilis they are not.

When it is said, the expression and the semantic units have the same marker, it is meant that spatial properties governing the generation of the expression unit and the corresponding sememe are the same, such as in the case of the iconic representation of the sun, where, the semantic property of the sun being spherical, and the corresponding graphical figurae have the same spatial feature of being circular. This is, as would be remembered from section 4.2.2.3 is called toposensitivity.

When some of the properties of the semantic model are *topological*, and the corresponding expression unit has the same *spatial properties* that of the sememe, its production is governed by rules of similitude[32] prescribed by cultural conventions[33].

The other important difference between iconic codes (governed by ratio difficilis) and codes governed by ratio facilis, as noted in 3.6.2 is that, elements of the graphic code does not have positional and oppositional value out of the context they are put in use.

To summarize the above two sections, it can be said that: *abstraction* governs the transformation from the perceptual model to the semantic model, *rules of similitude* govern the transformation from the semantic model to the expression units in the case of ratio difficilis, and *convention* (arbitrary coding) governs in the case of ratio facilis. Both abstraction and similitude rules are, furthermore, conditioned by *cultural* selections, values, and rules.

[32] Eco (1976, pp. 199-200) defines 'similitude' or 'similarity' in this context as a 'transformation'; i.e. *biunivocal* correspondence of the points in the effective space of the expression and the virtual space of a content model, which defines abstract semantic relations.

[33] A transformation is a matter of *conventionalizing* procedure, whereby only some properties deemed as pertinent by the code are preserved: "A transformation does not suggest the idea of natural correspondence; it is rather the consequence of rules and artifice" (Eco, 1976, p. 200).

4.3.2 The content model in IR

In this section, the above described tri-layered model is examined in the context of documentary information retrieval situation.

In section 4.2.3, in discussing the type/token ratio, the content model in IRS is briefly discussed. It is the purpose of this section to describe how the content plane in IRS is structured and how this results in production of signs (which are determined to be governed by ratio difficilis in 4.1.2.3).

To explicate the structure of the content plane in IRS, we need to consider the structure of the database (index language, matching function, so on), the documents and their relation to the socio-cultural systems (institutions, etc.) in which they are produced.

4.3.2.1 The Social text

In section 4.2.3, the concept of intertextuality has been introduced, in order to illustrate that documents, or texts in general, are not objects in the sense of physical objects. It was noted in the same section that, each text is a meeting place of several texts which refer one another. It was also said that, text should be understood as any interpretable system, not only verbal texts. As mentioned in 3.1 any object inserted in a signifying practice, i.e. taken as a standing for something else, is a sign, therefore an object of semiosis.

Peirce's idea of (unlimited) semiosis (3.5.2), makes it clear that, each sign is interpretable only in terms of other signs. This corresponds very closely to the idea of intertextuality. However, it is worthwhile to expand on intertextuality a little more than this to understand the social dimension of semiosis (i.e. signification process, the signifying practice).

Kristeva (in Moi, 1986, pp. 75-77) considers (not unlike many other semioticians, e.g. Barthes, 1985; 1972; 1967) any *social practice,* such as; economy, science, art, etc., as a signifying system, that can be analyzed in terms of semiotic categories[34], or language-models.

Intertextuality, should be understood in the above context. For Kristeva, intertextuality means "... transposition of one (or several) signifying sign-system(s) into another ..." (ibid, p. 111), and referring to work of Bakhtin, another important literary theorist, Kristeva agrees with him that (ibid, p. 37) "... *any* text is constructed as a mosaic of quotations; any text is the absorbtion and transformation of another" (my *emphasis*).

These sign-systems or texts, to re-iterate, are not only verbal systems, but all sorts of signifying practices, including the social ones. The totality of these signifying systems is what Kristeva calls the *social text* (ibid, p. 87).

The social text, as Kristeva defines it, is prior to any representation, it is the totality of human practices seen as productions and/or transformations. Defining them as productions, Kristeva emphasizes that they are not reducible to representation or meaning, therefore, are not measurable. These social productions are the background against which systems of representation

[34] Not necessarily those that have been developed so far. Kristeva makes clear that, she understands semiotics as development (production) of (new) models (in Moi, 1986, pp. 76-77).

are based upon.

Systems (codes) are structures which establish equivalences[35] between elements of expression and content planes. In this way they assign value to the units of the system. The production, in the sense Kristeva employs, is anterior to any value system, it is prior to the product (value), this is why it is unmeasurable. The social text as production, in this sense, is a field, a meaning *potential,* which does not represent or express, but has the potential of doing so at its disposal (ibid, pp. 74-88).

This is close to the idea of *writing* as production,,as proposed by Derrida. For Derrida (1974), writing is made of differences or differential traces ('gramme') which he designates with the term 'différance'. *Différance* is much more than the *difference,* as found in Sausure's version of semiotics (Kristeva, 1989, pp. 332-333). Derrida maintains that writing is 'perpetual *deferring*' (thus différance as; difference + deference) of meaning, which does not close up as a *stable* system (Norris, 1982, p. 29). Writing in the above sense is similar to the idea of production, as understood by Kristeva, as being anterior to representation and meaning (in Moi, 1986, p. 83).

The social text as described above is the background against which the semantic model in IR is based upon. The semantic model is part of a representation schema. This schema is based against the background of a non-schema, which is the totality of social practices, i.e. the social text. The social text as a background against which the semantic model in IR is built corresponds in this respect to the level of stimuli in Eco's model of transformations in production of signs (fig. 4.3). As it would be recalled from 4.3.1.1, stimuli is the background against which the semantic model via the intermediary of the perceptual model, in the general model of sign production of Eco, is generated.

I will call this layer of the transformative structure in IR against which the semantic model is built, 'the social text' after Kristeva (fig. 4.4).

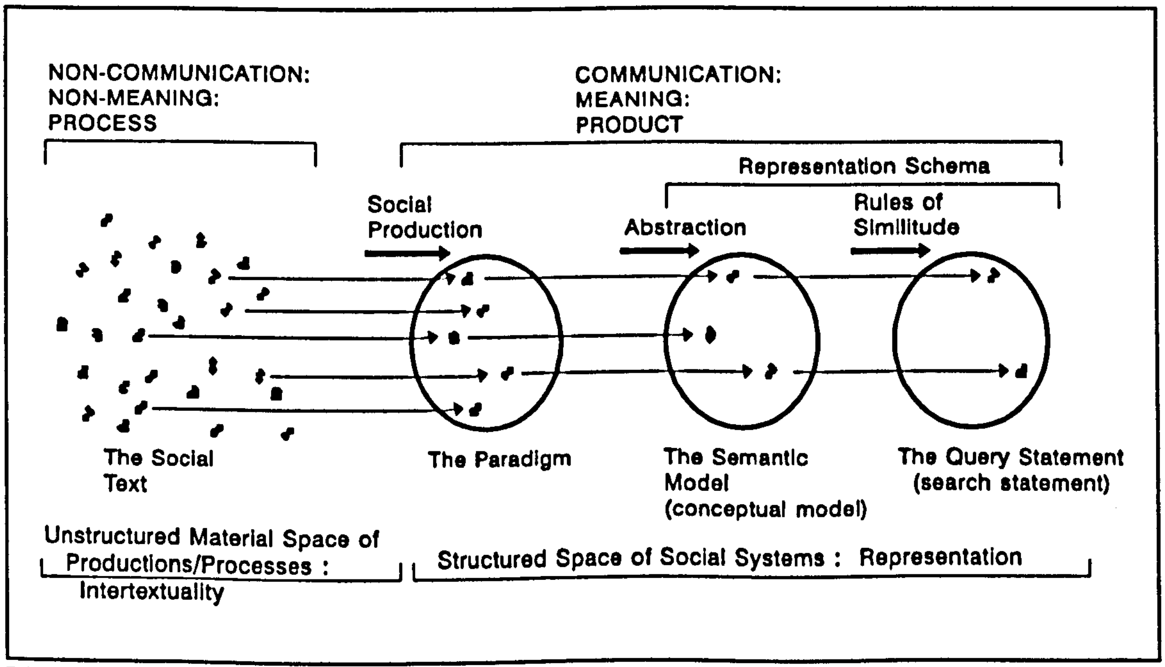

NON-COMMUNICATION:
NON-MEANING:
PROCESS
COMMUNICATION:
MEANING:
PRODUCT
Representation Schema
Social Production
Abstraction
Rules of Similitude
The Social Text
The Paradigm
The Semantic Model (conceptual model)
The Query Statement (search statement)
Unstructured Material Space of Productions/Processes : Intertextuality
Structured Space of Social Systems : Representation

Figure 4.4: The Tripartite Model of Sign Production in IR

[35] An alternative/complementary model is presented in 5.7.

4.3.2.2 Mapping from the social text

There are two levels of transformations that the social text undergoes before it is transformed into a semantic model in a documentary information retrieval system.

The first of these transformations involve reducing the multiplicity of meanings that potentially the social text can acquire to a smaller set of controlled (coded, valid, acceptable) meanings (corresponding to the perceptual model in fig. 4.3). At this level, meta-rules which allow interpretation of a text in such and such a way are *prescribed.* These prescriptive meta-rules are embodied in the institutions, systems, artifacts, professional associations, journals, papers, etc, which altogether constitute *paradigms* in science.

Ellis (1992), discussing paradigms in science, refers to Masterman, who identifies some 21 different senses of the term as used by Kuhn. These fall roughly in three categories:

1. metaphysical paradigms;
2. sociological paradigms;
3. artefact or construct paradigms;

of which, 'artefact or construct paradigms' are the more fundamental of the three categories and precede the other two.

Ellis also quotes Kuhn, who points out that: "A paradigm governs in the first instance not a subject matter but rather a group of practitioners" (ibid, p. 55). It is well known that these practitioners set up personal networks of contact comprised of small number of selected members known as the invisible colleges (Crane, 1972). In this broader sense of the term, the concept of paradigm should be understood in this section.

Lyotard (1984), in examining the condition of 'knowledge' in capitalist societies, underlines the role of consensus among experts in a field, in setting up prescriptive meta-rules which determine the validity of scientific statements. Scientific statements are *denotative statements* that describe objects and may be declared true or false. Scientific statements (knowledge) are further constrained by the following requirements: *i)* the objects to which they refer must be available for repeated access, *ii)* it must be possible to decide whether or not a given statement pertains to the language judged relevant by the experts[36] (ibid, p. 18).

A scientific statement, according to Lyotard, is a language game in the Wittgensteinian sense (c.f. 3.9) which affects the pragmatic (pragmatics in the sense of section 3.5.4) posts of sender, addressee, and referent in a certain way; the sender should speak the truth about the referent, i.e. should be able to refute any opposing or contradictory statements about the same referent; the referent which the sender speaks of, is supposed to be expressed by the statement, in conformity with what it actually is. This is what is known as verification or falsification (see, e.g. Popper in Miller, 1983, pp. 141-151). Lastly, it should be possible for the addressee to give or refuse his assent to the statement he hears. In other words the sender needs an equal partner, who can in turn become the sender. "The truth of the statement and the competence of its sender are thus subject to the collective approval of a group of persons who are competent on an equal basis" (Lyotard, 1984, p. 24). In short, the sender must formulate the statement according to the rules of the game and petition to the addressee to accept it.

[36] Thus, the 'consensual' nature of the scientific knowledge (see below).

However, this request for acceptance is not a denotative but a prescriptive language game. Therefore, the scientific statement which is denotative in nature, relies on a prescriptive one in order to legitimize itself (cf. 5.5.4). Prescriptive statements are consensual statements, and therefore by definition cannot be declared 'true or false'. "The argumentation required for a scientific statement to be accepted is thus subordinated to a "first" acceptance (which is in fact constantly renewed by virtue of the principle of recursion) of the rules defining the allowable means of argumentation" (ibid, p.43). The acceptability of scientific propositions is therefore, foremost depends on the contract drawn between the equal partners in the game. The *admisiblity* of scientific statements ultimately depends not on 'scientific knowledge', but on a totally different form of knowledge; 'the narrative knowledge' or 'narration', which has no extensional, i.e. denotative value (ibid).

The scientific discourse requires a formalized language, which involves definition of an axiomatic: "that includes the definition of symbols to be used in the proposed language, a description of the form expressions in the language must take in order to gain acceptance (well-formed expressions), and an enumeration of the operations that may be performed on the accepted expressions..." (ibid, p. 42).

Godel (1962) shows that, in any consistent formal system, such as 'the arithmetics', there exists certain propositions that can not be proven from within the system. This is why, as Lyotard states, the scientific language relies on a metalanguage (meta-rules), which are not provable, but subject to consensus between experts. "When a denotative statement is declared true there is a presupposition that the axiomatic system within which it is decidable and demonstrable has already been formulated, that is known to the interlocutors, and that they have accepted that it is as formally satisfactory as possible" (Lyotard, 1984, p. 43). The result of this non self-completeness is that, a new argument ('move', in language games terms) is either performed within the established rules, or, involves the invention of new rules, i.e. a change to a new game (ibid). This will be further discussed in 5.5 when rule-following and rule-changing activities are discussed in terms of language usage (sign production).

This level of meta-rules (or paradigm) which codes the social text into acceptable propositions is located beneath a more abstract level which organizes it into documented information, reducing the multitude of meanings at the level of paradigm yet into a smaller, more restricted (encoded) set.

4.3.2.3 Scientific information communication

It is generally suggested that scientific information is communicated among scientist through either formal or informal channels (Crane, 1972; Garvey, 1979).

Published articles in learned/professional journals (including conference precedings, etc.) are one of the more important formal communication channels in science. The entire system of publishing research through journals, relies on what is known as peer refereeing or evaluation (Garvey, 1979, pp. 69-70). The contractual nature of scientific knowledge has been discussed in the preceding section. It is suffice here to note that, the production of published scientific information is similarly governed by a set of *contractual* rules.

The scientific communication chain is further composed of secondary sources that index, abstract, etc. the published information. For the purpose of this section, it can be assumed that these publications are either abstracted or indexed as a full text in a database[37]. The documents represent the activities (practices) in a paradigm, mainly in terms of verbal language. They abstract the paradigm they are produced in[38], thus reducing its rich meaning potential to a controlled sub-set.

Documents in a database (of whether primary or secondary source), comprise the second level of transformations (fig. 4.4) that leads to the production of expression in IRS. They constitute the semantic model (type) to which the query terms are accorded (cf. 4.2.3). More accurately, the semantic model is comprised of representation of the documents in the database (see footnote 22), therefore includes the indexing rules (index language) and the inference rules (matching function)[39].

It should be however made clear that, the actual physical structure of the documents constitutes the model from which the expression units are generated. This is, as noted in 4.2.2.3, a topological condition, meaning that the form of the constituting units of the content plane (i.e. the spatial relation between them) governs the form of the units of the expression plane (cf. 4.1.2 and 4.2.3).

To demonstrate that document retrieval systems work by following (topological) similitude rules, it is sufficient to look at the matching process in IRS. In all IRS, whatever the actual form of the matching function is, it always works by a process known as pattern matching, which literally matches the physical patterns on the documents[40] with the physical patterns of the query terms. To retrieve documents it is thus necessary and sufficient to specify the patterns of graphic traces to be matched.

The following sections deal with the coding that establishes equivalence between the expression and the content planes. It will be shown in 4.5.2 that, in IR there are more than one code in play, establishing correlation between the expression plane and more than one content planes. The topological relation will then shown to be just one of the codes in documentary information retrieval situation.

4.4 Coding, Undercoding, Overcoding

When codes is discussed in 3.5.9 the distinction between an s-code and a code has been drawn. Code in the weak sense of the term means a correlational code. A correlational code establishes equivalence between an expression unit and a content-unit.

[37] These publications (full text or surrogates of them) are referred as documents in the jargon of IR and throughout this dissertation.

[38] Which is not solely composed of verbal texts, but all sorts of different signifying practices.

[39] One can envisage the case where the documents are stored and searched full-text, without an intermediary mechanism. This is a special case of the general method of storing/retrieving documents, however it does not affect our discussion.

[40] In most IRS, matching takes place via an intermediary mechanism in the form of an index, called the 'inverted file'. This is, of course, still a *pattern matching* operation.

Such is the case in verbal languages. An expression unit such as /tree/ is correlated to the content unit <<tree>>, which has a relatively precise meaning (i.e. its delimiters or markers are foreseeable; see footnote 31).

There are cases however, where, precise units of expression correspond to a discourse and concatenation of smaller expression units correspond to relatively compact content units. The former is known as 'undercoding' and the latter as 'overcoding' (Eco, 1976, pp. 129-142).

4.4.1 Undercoding

Undercoding happens when a content unit that can not be broken further into simpler (more analytic) units stands for a vast idea, or more accurately for a whole discourse (ibid, pp. 135-136).

Consider the case of screaming "fire!". The exclamation "fire!" does not only denote that there is a fire somewhere or something is burning, but also may mean "something or somewhere such and such is burning, and help is needed, call the fire brigade, etc.", depending on the contextual and circumstantial selections[41]. The expression-unit /fire/ which can not be further analyzed into simpler meaning-bearing units stands for a relatively complex discourse such as conveyed in the preceding sentence in quotation marks.

Another example is that of a painting, say portrait of a person. A sign of this kind which for most practical purposes can not be further analyzed into smaller meaningful constituent parts, is an expression unit. The content of such a portrait is not only person *X*, but stands for a whole complex discourse, such as; "middle aged woman with short white hair and delicate eyes, looking as if the long years she ... etc. etc."

4.4.2 Overcoding

Overcoding works in two ways: either; given a code assigning content to certain minimal expression units, another code assigns new content to larger chains of expression units of the previous code, or, it may work in the opposite direction; given a code establishing content to a certain expression units, another code analyzes these units into simpler units.

An example of the former is ready-made syntagms such as "how are you", "I beg you pardon", which work as minimal units with atomic meaning. The latter case happens when different pronunciations of a word correspond to a different shades of meaning (ibid, pp. 133-135).

[41] Eco (1976, p. 106) gives the following definition: "Contextual selections record other sememes (or group of sememes) *usually* associated with the sememe in question; circumstantial selections record other sign-vehicles (or group of sign-vehicles) belonging to different semiotic systems, or objects and events taken as ostensive signs, *usually* occurring along with the sign-vehicle corresponding to the sememe in question ...".

4.4.3 Equivalence or inference

When discussing codes in 3.5.9, the distinction between weaker and stronger senses of the term has been introduced. The weak sense as indicated there, means correlational codes, which establish equivalence between the units of the two planes.

It is however rarely the case that correlation (equivalence) is not accompanied by discursive rules, which trigger inferential processes.

A code, such as the 'Morse alphabet', is an example of the weak sense of code, which establishes one to one equivalence between a set of expression devices and the letters of a natural language, say, the English alphabet.

In many other cases however, the equivalence relation is supplemented and complicated by presence of instructions which make possible interpretation of expressions.

The discursive rules or instructions make possible contextual selections, thus complicating the initial equivalence with set of possible interpretations (see section 5.6.1.2 for the discussion of code as an inferential mechanism).

Eco (1984, pp. 164-188) shows that, even s-codes that are not suppose to have any signification power (3.5.9), just establishing structure to only one of the semiotic planes, do sometimes by virtue of its internal rules that govern the functioning of the system produce signification. Thus, concludes Eco, it is virtually not possible in reality, to distinguish between, s-codes, correlational codes, and codes with inferential rules, etc. This is the 'strong' version of the term code (ibid, p. 182) that has to be borne in mind, in discussing codes in IR context (cf. 5.6.1.3).

4.5 Coding in IRS

This section describes the structure of IRS which is taken as a code, that is a semiotic system. The purpose of the following paragraphs is to explicate the general structure of coding in IRS.

4.5.1 The type of coding in IRS

In IR, the matching function establishes equivalence between the query terms and the documents in the database. The query (in Boolean systems) can be shown to be equivalent to:

$$t_1 \; OR \; t_2 \; OR \; ... \; OR \; t_r$$

where; *OR* is the usual *disjunction* operator of boolean systems, *t*'s are either simple *conjunctions* of index terms or their negatives; $t = i_1$ AND i_2 AND ... AND i_n, where; i_j denotes either '*i*' or 'NOT *i*', for *i* one of the *n* index terms (Bookstein, 1989, p. 469).

The corresponding content to this, is a set of texts (either full-text or surrogates of them), comprising a discourse (cf. 4.2.3). It is evident from this simple analysis that, the type of coding in IR is that of undercoding[42]. The main function of IR, thus emerges, as establishing equivalence (correlation) between relatively small expression units and larger units of content, that is text. This is a typical case of (gross) undercoding.

4.5.2 The two levels of coding in IRS

It has been mentioned briefly in 4.3.2.3 that the first stage in document retrieval is to describe the form (physical layout), i.e., the spatial disposition of the documents that is to be retrieved.

It has shown in 4.1.1 that, the commutation test reveals when applied to IRS, the pertinent units of the expression plane are those of the 'query terms' and that of the content plane are the retrieved documents. This demonstrates that, IRS at its most basic level, or more accurately as an initial code, correlates (cf. 5.6.1.3) the query with a set of documents whose physical form is isomorphic with that of the query statement.

At this level the meaning of expression terms (the query statement) is therefore the retrieved documents. Query statement means no more than, "find all the documents whose form matches that of the query statement", at this stage. The meaning of a query is thus, an abstract description, a schema, which denotes the form of the requested documents[43]. This is overcoding, in that, an additional meaning which resembles sememe (cf. 3.6.1.4) is attached to the expression units[44] (cf. 5.4.1).

The query stands for a set of documents whose form is set out by the query statement. When some documents are retrieved and presented to the enquirer however, the documents themselves stand for something else[45]. There is therefore, at least another code in play, which replaces the previous one as soon as the documents are retrieved, which relies on it as a anterior step (cf. 5.4.2). This is a typical case of metonymic substitution (in the sense of section 3.8), an initial set of expression units being replaced by another set of expression-vehicles (*fig 4.5b*). As it would be recalled from the discussion in section 3.5.13 that, this is the 'connotative' structure in the form of (ERC) R C (it will be referred as the ERC model, from now on).

[42] Each index term is coupled with one or more documents. A document, in a documentary information retrieval system is a text, corresponding to discourse in the content level. See also, 5.6.1.

[43] In 5.7.3, it will shown that an alternative description in terms of the concepts of sememes and interpretants is possible.

[44] However, except the rare case of where one wants to retrieve all documents containing certain terms, this can hardly be considered the 'real' meaning of the query. Therefore, 'overcoding', in actual *pragmatics* of IR is replaced by undercoding (see below, also 5.4.1).

[45] This is undercoding, query standing for a discourse.

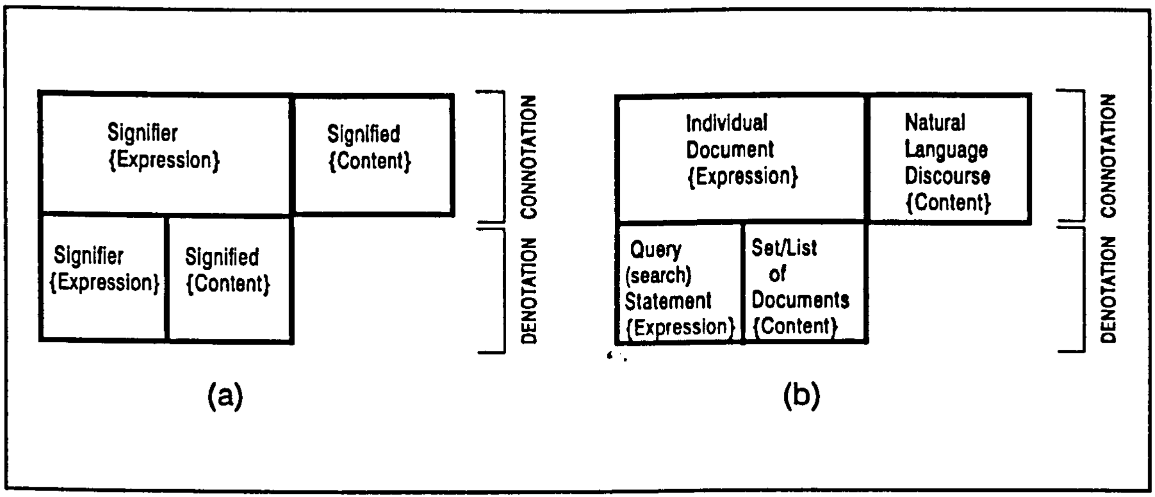


Figure 4.5: Denotation/Connotation in IRS

The initial level which is called denotation in the ERC model, establishes correlation between the query and the documents that are retrieved by it, which as a whole constitute a discourse in a field. The query statement is an abstract representation of this discourse, a type or model which selects the pertinent features of it.

The connotative code which relies on this denotative level, establishes another type of system of relations, this time that of natural language (or of a register[46]). By this time, the equivalence relation is complicated by inferential rules[47] (figure 4.6; cf. 5.6.1.3).

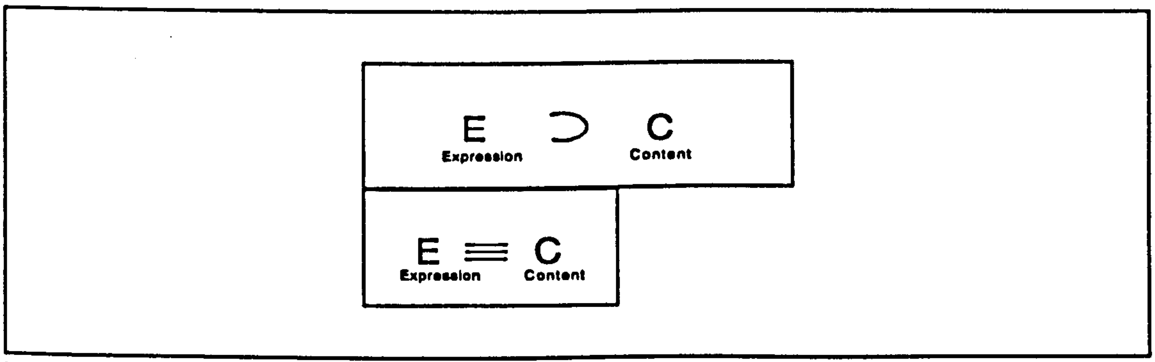


Figure 4.6: Infemtial Nature of IR (after inferential nature of connotation given in Eco, 1984, p. 34)

The consequences of this two level of coding in terms of information retrieval practices are examined in section 5.6.

[46] A register is defined by Halliday as "... what you are speaking, determined by what you are doing, and expressing the diversity of social process ..." (in Karlgen, 1993. p. 349). In this sense, *a register* is not simply a sublanguage (ibid).

[47] At this point, it may prove to be more fruitful to abandon the correlational model for a model based on 'networks' of relations (i.e. IRS as an *encyclopedia*; cf. 5.7.3).

4.5.3 IRS as a code

This section attempts to explain how IR as a particular code works, that is, the two level coding system described in the preceding section work together as a single system. It will be showed that although it is necessary in order to understand the structure of IR to separate the codes that make up the system, in practice it is hardly possible to draw a clear cut line between the two codes.

First, it is necessary to examine the case of denotation in more detail. The denotative level in IR codes the expression (query statement), which is made up of recognizable (replicable), discreet, meaning-pertaining units (individual words/terms in the statement) that can, in most cases, be further broken down into smaller non-meaning carrying elements[48].

The corresponding content however, is not a simple atomic unit, which can be easily recognized (or replicated). It is too made up of smaller discrete units (which is called conveniently as documents), however, each single document is not easily analyzable into its constituting parts[49], therefore not replicable (i.e., one can not reproduce the expression from its content). As a whole they can be represented by some terms/descriptors, which describe them as a schema. Only at this level of abstraction one can replicate the expression by analyzing[50] its content (which is a set/list[51] of documents). At this highly abstract level one can thus analyze a query statement, say, in a foreign language database, in terms of its effect[52] on the retrieved set/list. The expression in IR is, thus, only replicable at the denotative level. At the connotative level one deals with a natural language discourse (cf. 4.5.2). Each document is open to interpretation and re-interpretation, and in most cases impossible to be represented by another smaller, more precise (less ambiguous) set of signs, such as the propositional calculus used in logic (cf. 3.9 for Wittgenstein's position on this). Although two identical query statements describe the same set of documents, interpretation of each document and the use they are put in varies across the users[53]. This marks the major gap between the two modes of coding in IR, namely; ratio

[48] It is therefore, produced as a result of first and double articulations (cf. 3.6.1 and 3.6.2).

[49] One can of course analyze a document into its constituting features such as, stems, words, phrases, paragraphs, etc; however, cannot backproject from a single document to the original query statement (and hence the information need) that might have retrieved it in the first instance (see footnote 53).

[50] Using a technique, such as the *commutation test* (cf. 3.6.1).

[51] No differentiation is made between the 'set-oriented' view (e.g. Bookstein, 1989) and the 'document by document' perspective (e.g. Robertson, 1977a), as the argument applies equally well to the both cases.

[52] Such as, the size of the retrieved set (or the positions of the documents in a ranked list), the *topological* characteristics of the retrieved documents (i.e. the combinational characteristics of the query terms), so on. In other words, only in terms of the physical, spatial features of the documents retrieved.

[53] This is a well known phenomenon in IR, two users with identical query statements tend to almost always have different ideas regarding the relevance of the retrieved documents.

difficilis (iconic[54]) and ratio facilis (verbal). Yet, this very same gap makes possible to work IRS as signifying systems. Taken alone, denotative code works as an abstract schema without any apparent linguistic meaning. However, together with the next level of coding, the connotative code, the overall assemblage constitutes a language. The gap defined above is somehow bridged at the moment of human semiotic act of production, and results in semiosis (cf. 5.4.1). Yet again, the gap marking the two modes of coding is much more than trivial.

Kristeva's comparison of the characteristics of signs and symbols illuminates the differences between the denotative and connotative codes in IR. From her discussion of the symbol; "the quantitative limitation, repetition, and general nature of the symbols" (in Moi, 1986, p. 65), emerge as the characteristics of the symbolic semiotic practice.

The symbol is general in its representation capacity, corresponding to text/discourse; evoking a collection of associated images and ideas, rather than smaller units of meaning. The relation of the symbol to its object (referent, symbolized) is that one of restriction, that is, the symbolized being a universal transcendence, cannot be reduced to the units evoking it (the symbol) (cf. 5.5.5). The symbol is restricted in its ability to be articulated horizontally (i.e. in relation to other symbols). It is the characteristics of the symbolic mode of production to be repetitive, that is, same patterns of symbols are generated indefinitely (ibid, pp. 64-70).

The sign on the other hand is characterized by its ability to deflect progressively in everlonger chains of syntagms that creates the illusion of an open structure, an open system of transformations. In contrast to the symbol, the sign is boundless in its capacity to be articulated horizontally. Its meaning is a result of interaction with other signs, and in this respect, has the capacity of generating and transforming new structures. The units of the sign system, the words, the lexemes, are much more concretized than the symbol, referring to entities of lesser dimension (ibid, pp. 70-72).

The above discussion makes clear the differences between the two very different types of coding that constitute the denotative and connotative codes in IRS. However, the consequences of this, in terms of 'productive labour' required on the part of the users of IRS, require a much more detailed analysis which is the subject of the next chapter.

[54] Ratio difficilis which governs the production of expression (query statement) in IR governs also the production of signs which are commonly referred as, icons in Peirce's sense (cf. 3.5.3) and symbols in Saussure's terminology.

Chapter 5
A Theory of Sign Production in IR

The preceding chapter has dealt with the structure of IRS viewed from a semiotic perspective. This chapter deals with the actual practices (labour) which give rise to production of signs (expression or query statement) in IR.

Whereas, a theory of codes according to Eco (1976, pp. 152-153) concerns "... both with the structure of sign-function and with the general possibility of coding and decoding", a theory of sign production concerns with (ibid); "... the kind of labour required in order to produce and interpret signs, messages, or texts (physical and psychological effort in manipulating signals, in considering, or disregarding, the existing codes; time needed, degree of social acceptance or refusal, energy expended in comparing signs to actual events; pressure exerted by the sender on the addressee, and so on)".

In dealing with the production of signs in IR situation we will go back to the questions set at the beginning of the chapter 4, in particular to the relation between the query and the query statement, the retrieved documents and the idea of relevance, and attempt to answer at least some of them.

5.1 A classification of types of labour in producing signs

Eco (1976) gives a typology of labour required to produce signs. This involves four categories of physical labour, namely; *recognition, ostention, replica* and *invention* (see fig. 5.1).

Recognition occurs "when a given object or event, produced by nature or human action (intentionally or unintentionally), and existing in a world of facts as a fact among fact, comes to be viewed by an addressee as the expression of a given content, either through a pre-existing and coded correlation or through the positing of a possible correlation by its addressee" (p. 221).

Ostension occurs "when a given object or a event produced by nature or human action (intentionally or unintentionally and existing in a world of facts as a fact among facts) is 'picked up' by someone and shown as the expression of the class of which it is a member" (pp. 224-225).

Replica governs the production of the most usual types of expressions such as, 'phonemes', 'morphemes' and 'lexemes' (c.f. 3.6.1) that are produced according to 'ratio facilis', as well as, 'vectors' which are subject to ratio difficilis, and stylizations lie somewhere between the two types of ratios (pp. 227-228).

Invention is different from the other three modes of production in that, there exists no previous convention that correlates the elements of the two semiotic planes. Invention is a mode of production "... whereby the producer of the sign-function chooses a new material continuum not yet segmented for that purpose and proposes a new way of organizing (or giving form to) it in order to map within it the formal pertinent element of a content type" (p. 245).

Whereas in recognition, ostention, and replica, the correspondence between a token and its type (whether ratio facilis or difficilis) is known (accept in the case of positing for the first time a correlation between a token and its content in the recognition of imprints, symptoms, and clues; cf. 5.3.1) as the type exists as a cultural product, in inventions the sign producer must somehow posit this relation.

5.2 A Typology of Modes of Sign Production

Eco (1976, pp. 217-257) proposes a classification of different ways of producing signs, based on the type of labour involved in their generation.

The categories of modes of production are determined according to four criteria; the physical labour involved, the type/token ratio, the physical continuum of the expression plane, and the mode and rate of articulation (see fig. 5.1).

The criterion of type/token ratio (cf. 4.2.2) and the type of physical labour involved in production of signs (cf. 5.1) have been discussed in the previous sections. The modes of sign production are also characterised by the physical medium shaped by the expression form in producing the expression functives. The medium shaped by the expression form is classified into two categories, either, *i)* the expression and its *possible* referent is made up of (shaped) by the same material, that is homomaterial, or *ii)* they can be made of different material, i.e. heteromaterial. In the case of being heteromaterial, the material that shapes the expression and its potential referent, is either *1-a)* arbitrarily selected or *1-b)* in a few cases, imposed by the direct action of its referent.

It is worthwhile to note that, in the above categorization only a possible or potential referent is considered. It is not important whether such a referent does really exit or not since the idea of an external referent is not a postulate of the semiotics that has been adopted for the purpose of this dissertation (3.5.14). It is not necessary for the sort of semiotics adopted here to refer to some external referent that sits 'out there' objectively. It is however useful, especially in the cases of recognition and ostension to consider the relation between the sign-vehicle and its possible referent when they are used for mentioning things. This is of no particular use for the purpose of this dissertation and only discussed for the sake of completeness of the discussion of fig. 5.1. It will not be pursued further in the rest of the chapter.

Signs differ also with respect to the complexity of articulation of their combinational units; on the one end there are codes that are analyzed into precise combinational units and dully coded or overcoded, on the other extreme those whose combinational units are not further analyzed into smaller units thus, undercoded.

Figure 5.1 (taken from Eco, 1976, p.218), summarizes the modes of sign production that Eco proposes according to the above discussed four criteria. A brief discussion of the each category in this table is given below.

PHYSICAL LABOR required to produce expressions	RECOGNITION			OSTENSION			REPLICA					INVENTION	
TYPE/TOKEN RATIO: RATIO DIFFICILIS	IMPRINTS						VECTORS					CONGRUENCES	TRANSFORMATIONS
				EXAMPLES	SAMPLES	FICTIVE SAMPLES		STYLIZATIONS			PROGRAMMED STIMULI	PROJECTIONS GRAPHS	
RATIO FACILIS		SYMPTOMS	CLUES						COMBINATIONAL UNITS	PSEUDO-COMBINATIONAL UNITS			
CONTINUUM TO BE SHAPED	HETEROMATERIAL (MOTIVATED)			HOMOMATERIAL			HETEROMATERIAL (ARBITRARILY SELECTED)						
MODE AND RATE OF ARTICULATION	PRE-ESTABLISHED (coded and overcode) GRAMMATICAL UNITS (according to different modes of pertinence)											Proposed undercoded TEXTS	

Figure 5.1: A Typology of Modes of Sign Production (in Eco, 1976, p. 218)

5.3 The Modes of Sign Production

This section summarizes the characteristics of the individual modes involved in the production of signs given in Eco's schema (fig. 5.1).

5.3.1 Recognition of Imprints, Symptoms and Clues

In recognition of imprints, a ready-made expression is correlated to its content of class of all possible imprinters. Imprints are conventionally coded by a previous act of a series of mentions, inferences, etc. In order to be recognized, imprints first need to be correlated to their content by a process similar to ostention, replica, or invention. Once a cause-effect relation is established by one of these processes, recognition of the imprint occurs by backward projection from the expression (imprint) to its possible causes.

Imprints are doubly motivated, first by the presupposed relationship to their cause, and then again by the form of their content. Therefore, the type/token ratio in this mode is difficilis. The stuff the expression and its possible referent are made of is necessarily heterogeneous (motivated heteromaterial).

Symptoms are also correlated to its content by a previous act of coding, thus the expression is ready-made, however its markers do not have the same form (spatial properties) of its cause (content), therefore the type/token ratio is facilis (smoke does not have the same form as fire or red spots as measles). The expression stuff is dissimilar to its possible cause.

Clues are objects or traces which are not imprints, that are recognized to belong a precise class of agents. However, they are seldom coded, therefore, usually interpreted as a result of a complex act of inference rather than recognized. The type/token ratio is facilis and the expression continuum and its possible referent (agent) are made of heteromaterial.

5.3.2 Ostension: Examples, Samples and Fictive samples

In ostension, expression form is determined by the form of the content whose sememic composition is governed by the shape of the object used in the ostensive production. The type/token ratio is therefore difficilis. The objects used in ostension, however, are already produced in practice as functional objects and constitute a sort of repertoire, an expression system (s-code), therefore should also be considered as ratio facilis. Ostension is in this respect a particular category where both types coincide and becomes token/token-ratio.

When an object is selected to represent its class, it is an example. A cigarette can be picked and shown to mean "cigarettes" or "please buy some cigarettes". In the second case the object stands not only for its class but for an entire discourse.

If only part of the object is selected to express the entire object and thus its class, this constitutes a choice of sample. An example of the case of sample is when a musical quotation given to mean the whole composition, or a piece of fabric shown to refer to the entire cut, or indeed directly to the "jacket" itself made with the fabric.

Fictive samples are samples stand for its class, however not so much picked up as re-made to represent a particular act, gesture or sound. Such is the case when one mimics a particular action to mean for instance, "I punch you".

5.3.3 Replica: Vectors, Stylizations, Combinational units, Programmed Stimuli

In order a sign to be replicated, it should be analyzable into its constituting elements of more analytic type. Such is the case with natural language codes. A given syntagm (a phrase) is analyzable into elements of the first articulation, words, which are analyzable into phonemes or letters, that is the elements of the second articulation (cf. 3.6.1). In other words, pertinent elements of a given expression system combine with other pertinent elements of the same system to compose the expression units or functives.

Under the category of combinational units we have therefore the most usual expression units; phonemes, lexemes, morphemes, etc.

However, the verbal language is not the only code which is made of combinatorial elements. There are codes of zero, one, two, or non-fixed number of articulation (cf. 3.6.2), which are made of analyzable units that can be replicated. Naval flags, traffic lights, playing cards, tonal music are some examples of the variety of codes with different number of articulations. All these codes have the following characteristics in common: they are governed by ratio facilis, correlated arbitrarily (by convention) to their content and the expression continuum is arbitrarily selected (therefore, heteromaterial).

There are other codes however, whose elements do not combine with the elements from the same system but from other system(s) to make up an expression; these are called vectors.

The typical example of a vector is a pointing finger. A pointing finger is an example of kinesic pointers. These are characterized by having certain spatial features realized by a part of the human body conveying certain semantic features such as, closeness, direction, etc. The directional features in the case of a pointing finger orientate the addressee to "left", "right", "up", "down", etc. However, these directional features do not constitute an oppositional system in the form "left vs. right", "up vs. down", that can be used as combinational units in other kinesic configurations. They are used in conjunction with another system, such as in speech with the verbal language, to mean "to the left", "to the right", and so on. The pointers of this sort is governed by ratio difficilis since the pointed direction (the expression) and the implied direction (the content) is the same. The expression material is arbitrarily selected.

Other type of vectors include: spatia-temporal vectorialization in phrases such as /John beats Mary/ where the sequence determines the content such that the reversal makes the contents reversed as well; change in the pitch of voice during speech may cause change in meaning, such as humming with a upward pitch-curve might mean "questioning", etc.

There are other replicable signs which must not necessarily be combined with other features from any other system. These are called stylizations. Stylizations work by sort of overcoding (cf. 4.4.2), where expressions are produced by their similarity to already existing sign-vehicles which act as expression-types. Only some features, those sufficient to make the expression resemble its type, are retained, while a lot of free variation is allowed. Such is the case with the playing cards for instance where the images on the cards retain some of the features of their type, while various stylizations of the types are produced with different makes of cards. It is also the case with popular depiction of historical/mythical figures such as the Devil, Virgin Mary, etc.

Stylizations are subject to both ratio difficilis and facilis, since although produced according to the shape (form) of their sememe (such as a King on the playing cards denote the type King) thus obeying ratio difficilis by virtue of a sort of catachresis whereby previous inventions become expression-types from which replicas are produced by ratio facilis. This is an example of ratio difficilis transforming by repeated exposure to communication and successive conventions into ratio facilis.

There is yet another category of productive operations which seemingly obey simultaneously both to ratio facilis and difficilis.

There are cases where seemingly non-significant, non-meaning bearing elements, i.e. stimuli (cf. 3.5.3) are used for a semiotic effect. This sort of stimuli which act as an expression of some (partially) foreseen effect are called programmed stimuli.

When a stimulus such as a change in the tone of the voice (or any other 'para-linguistic' feature) inserted in a speech, or a flash of light in a theatrical performance with an aim to elicit certain response from the addressee, it should be considered a part of the semiotic competence of the sender.

In such a situation the sender wants to elicit certain behavioral or cognitive reaction from the addressee, therefore some of the effects of the stimulus are foreseen by the sender, however not all effects are predictable. This could be the case when a speaker elaborating a discourse in a judiciary rhetoric and trying to arouse in the addressee the feeling of sympathy by using certain para-linguistic devices. However, the producer of the expression does not know exactly how these devices will be received and interpreted by the addressee, so much that he/she is actually making or inventing than performing a programmed stimuli.

The expression is made of analyzable and replicable units (governed apparently by ratio facilis) produces a vague response, corresponding to a discourse on the content plane. The stimulus being generated in expectation of a certain reaction from the addressee is at the same time governed by ratio difficilis. Since, not all the effects of the stimulus are foreseeable the productive act lies somewhere between replicas, and inventions, proposing a new (tentative) coding. In this capacity, the programmed stimulus should be considered as a sign function, where the stimulus acts as the expression of a supposed effect which functions as its content plane. This type of productive operation will prove to be of particular importance in discussing the productive labour in the context of IR situation.

Systems with pseudo-combinational units are those expression systems with detachable, analyzable, features somehow appearing to be without any content. The content plane in such systems remains uncoded as it were open to all comers, as it is the case with most abstract painting and atonal music.

5.3.4 Invention: Graphs, Projections, Congruences

In inventions there is no pre-established code correlating pertinent features of the expression system with the pertinent elements of the content plane. In this respect, it is radically different from all other modes of production discussed above. The producer of expression in inventions needs to give form to a heteromaterial expression continuum and posits a correlation that maps the features of the expression with that of the content. This mode of production is thus, governed by ratio difficilis.

In order this proposed (posited) correlation to be successful, i.e. recognized and accepted culturally, the addressee of the message (expression) should be able to successfully map from the expression to the underlying conceptual (semantic) model (cf. 4.3.1). Inventions therefore should be understood as acts of code making. There are two types of code making; 'moderate inventions' and 'radical inventions'.

A moderate invention occurs when the producer of the expression maps directly from a perceptual model (cf. 4.3.1) to an expression-continuum and establishes an expression form which dictates the rules to produce the corresponding content model (fig. 5.2). This is a different procedure than the three layer model discussed in chapter 4 (cf. fig. 4.2). Many artistic productions, including the classical paintings are this type of invention.

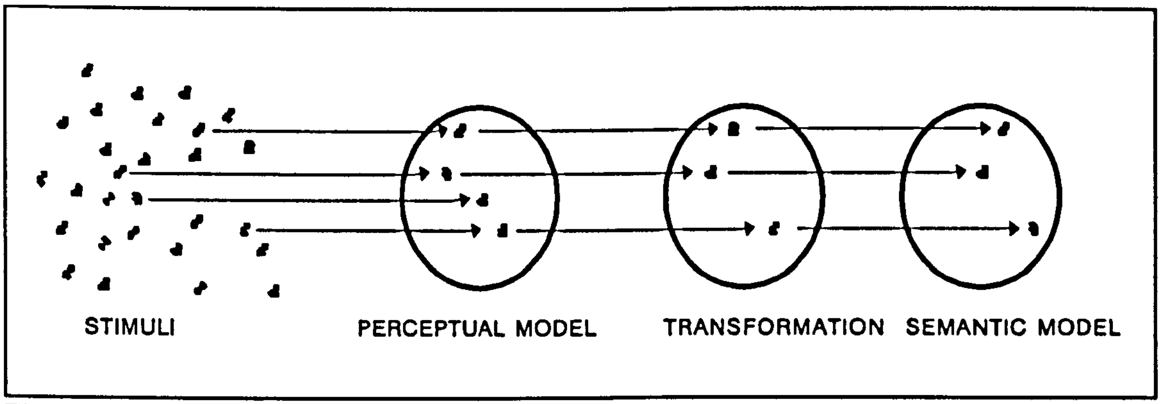


Figure 5.2: Moderate Inventions (in Eco, 1976, p. 253)

A radical invention occurs when the sign producer bypasses the existing perceptual model and maps directly from the stimuli into an yet unshaped expression continuum (fig. 5.3). The addressee who receives the invented sign-vehicle not only has to establish the content model corresponding to it, but must first construct the underlying perceptual model by a process of successive trials of guessing. This is a case a violent new proposal which upsets the previous conventions (see the discussion of rule governed and rule changing activities in section 5.5).

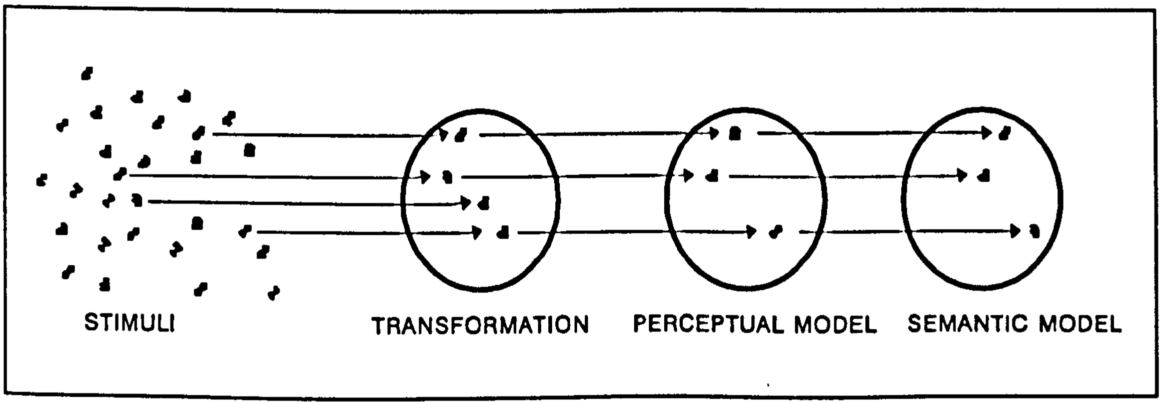


Figure 5.3: Radical Inventions (in Eco, 1976, p. 255)

In radical inventions it is very likely that the proposed code does not establish at all or succeeds to get acceptance only after a long period of struggle against rejection and subjugation, as in the case of all great artistic innovations as well as major scientific paradigm shifts and changes.

There are three major categories under the heading of inventions; *graphs, projections*, and *congruences*.

Graphs are topological transformations in which spatial points in an expression space maps onto points of a non-toposensitive relation on the content plane. Such is the case of Peirce's existential graphs in which spatial relations in the graphic expression correspond to non-spatial relations on the content plane. To illustrate this consider the expression: "every dependent worker belongs to the class of exploited and alienated proletarians", this can be represented graphically as in fig. 5.4 (in Eco, 1976, p. 258).

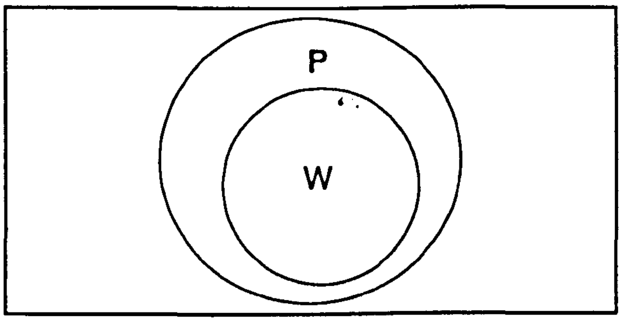


Figure 5.4: Peirce's Existential Graphs

Projections are topological transformations in which spatial points in an expression space maps onto points of a toposensitive relation on the content plane. There are strong similitude rules at play in projections which map pertinent features of the content plane onto the expression continuum without changing the spatial disposition or topological properties of the content (semantic) model. Thus they are highly *toposensitive* (ruled by ratio difficilis) as indicated its relative vertical position in figure 5.1. However, the similarity between the content model and the expression form is not always self evident, and the rules of the transformation must be learned (see the criticism of iconism in 4.2.1). The correlation between the two planes is thus conventional as usual, hence, it is always possible to map backward from the expression to the content model which does not exist (extensionally null set), as in the case of portraits of mythological figures in the classical painting.

Congruences or casts are point to point correspondences between the physical space of the expression and the physical space of a real object. An example of this is the death mask found in some cultures. They map the physical appearance of a person by means of conventions of similitude which keep only some of the physical properties of the human face, discarding the many others. These heteromaterial masks can of course be faked, corresponding to no actual person, therefore extensionally null. This is a case of extreme toposensitivity and the type/token ratio is full difficilis.

5.4 The denotative and connotative codes in IRS

As argued in 4.5.2, what is called the connotative level, which deals with the meaning of the documents (i.e., the level of natural language discourse), relies on an anterior code referred as the denotative level.

It is also pointed out in the same section that the denotative level itself does not deal with natural language as such. It is rather closer to what is known as the iconic code, whose form on the expression plane is similar to the form of its semantic model (i.e. the case of ratio difficilis).

This section investigates the process of formulation of query by the user of an IRS in an interactive situation and submition of it to the system for retrieval. For this purpose, the labour involved in the production of signs discussed above are examined below in the context of IR and the denotative and connotative codes.

5.4.1 The denotation

The initial step in retrieving documents in computerized systems involves the description of the spatial properties (i.e. topological relations) of the terms that are expected to be found in the documents that are to be retrieved (cf. 4.2.2.3). This is, as discussed before, an activity in which the expression form is produced according to the form of the content plane, or more accurately, the form of the semantic model which the expression aims to describe topologically. This sort of production is said to be ruled by a particular type of type-to-token ratio called difficilis.

The type-to-token ratio is *difficilis* (literal translation is *difficult*) in IR, that is, in the initial stage of formulating the query the expression-type to replicate the expression-token[1] from does not exists. This means that there is no expression system (i.e. expression s-code, cf. 3.5.9) in IR similar to, say, the phonological code which organizes the units of the systems into paradigms of mutually exclusive oppositional structures in the sense that Saussure used the term paradigm (cf. 3.5.7).

It can be however argued that there is such a system in IR after all. It is without doubt that the search terms are organized into paradigms either in the mind of the user or indeed in classification schemes and thesauri. More importantly, the index (or indexing language) of the system, as a matter of fact, organizes the terms in the documents into such a system, which certainly resembles an s-code. The major difference between this and a proper expression system such as the phonological code however is that, the s-code in phonemics is independent from any content system that it may be correlated to (i.e. arbitrary coding), whereas in IR, it is not. The expression terms in IR are motivated by their meaning/content. More accurately, the expression-type and the content-type in IR coincide. They are one and same. The expression-token is accorded directly to the content-type.

In discussing the kinesic pointers (5.3.3), it has pointed out that, they too are governed by ratio difficilis. However, there is one major difference between the two productive operations in that, whereas the individual expressions in the kinesic pointers, such as pointing my finger towards an object, is correlated with a precise meaning[2] (e.g. "this"), in IR the expression units are correlated to texts which are not precise atomic meaning units, but complicated discourse (see 4.2.3). This, as it would be remembered from 4.5.1, is a case of undercoding. In this respect, the productive labour in IR resembles more to programmed stimuli, where the expression is similarly made up of discrete, analyzable units that correspond to a vague discourse, a nebula-like behavioral or conceptual response.

Whereas kinesic pointers are easily replicable, programmed stimuli and expressions in IR are not (cf. 4.5.3). This is not because their expression is a dense textual-cluster which is not easily detachable and analyzable. Although the expression units are easy to analyze, the corresponding content, being a discourse and not an atomic meaning unit, even expressed, cannot easily be analyzed and recorded by its interpretants.

[1] That is, the search statement.

[2] This is why they are at once subject to both ratio difficilis and facilis (see fig. 5.1)

When one moves from this initial level to one step up in the hierarchy, to the connotative level, meaning in terms of the retrieved set is not valid any more, for now one deals with a different code, the natural language discourse. At the level of the natural language discourse, it is no more possible to determine precisely the relation between the original query terms and the text that one is reading, i.e. interpreting.

Taken individually, the initial level seems to be reproducible and it is like following the established rules which correlates the expression (query) with the retrieved set. Taken together with the next level of coding however, it is not reproducible. It can be viewed as proposing a new correlation between the expression (the query) and the natural language meaning of the documents, one can say for the moment, with the relevant documents[3], albeit the one which will never be recorded/accepted culturally for the reasons explained in the next section. Seen from this perspective, the productive operations in IR are more like rule changing or proposing new rules of interpretation (correlation), similar to the inventive labour discussed in 5.3.4. It is therefore worthwhile to compare the labour involved in IR with that involved in inventions.

5.4.2 The Connotation

In inventions, from the point of view of its producer/sender, the expression is a coded semantic model. The producer of the sign posits a relation between the expression and some content that has not been defined before (or at least, not culturally). From the point of view of the addressee of this yet culturally unknown expression however, it is just an expressive structure without an apparent content. It takes some interpretative labour on the part of the addressee to re-constitute the original relation defined between the expression and its semantic model which is until now only known to its producer. The addressee needs to map backwards from the expression to the original content model proposed for the first time by its generator (Eco, 1976, p. 252-255).

When this labour succeeds, that is, the proposed relation is recognized culturally, the proposal becomes an established code. If for some reason no interpretant could be found for this correlation, it fails to become a code (ibid).

When the inventive labour of code making succeeds in establishing itself as a new sign function, a new code, it becomes a source of material to be manipulated and transformed for new sign-functions, so the *semiosic* spiral starts to roll once again. If it is a relatively stable code, it can lead to the processes similar to stylizations where, by virtue of successive overcoding it becomes an expression-type from which other copies are replicated. The undercoded sign-function becomes an overcoded ready-made object. It becomes part of the general repertoire of signs that generates cultural habits, expectations and so on (ibid).

The difference between the situation in IR and this sort of code making is that, in IR it is practically impossible to restrict the semantic model into a definitive form. As discussed in 4.2.3, the documents which constitute the semantic model in IR, open up to different texts at various levels in a process called intertextuality or, semiosis.

Not only, there is no definitive interpretation of any given text, in other words the semiotic triad referred in 3.5.1 never closes, but any valid interpretation at a given moment will become obsolete in a later point in time. This will be discussed further in 5.5.5. For the above reasons,

[3] The idea of 'relevance' will be examined critically in 5.5.5 and 5.6.2.3.

there is no single way to establish a code which correlates a query statement with the natural language discourse. Therefore, the second level of coding in IR, called connotation, remains a sort of undercoding and fails to establish itself. It does not become an expression-type by means of overcoding from which others can be replicated. The relation between the query terms and the relevant[4] documents is that of a undercoding. For this reason, it is virtually impossible to foresee (predict) completely the relevant set of documents for a given query statement. If it was possible, this would be a result of a sort of overcoding, and the anterior stage of coding in IR (i.e. the denotative code discussed in section 5.4.1) would not be needed.

5.5 Rule-governed and rule-changing activities in IR

To fully appreciate the complexity of sign production in IR, it is necessary to focus on two important kinds of activity that one comes across in all sorts of different media; the rule-governed and rule-changing productions.

The rule-governed productions are those productive operations where, one produces an expression by mapping into it a new but foreseeable content. The expression is correlated to a content which is articulated by combining the existing (known) content-types. Such is the case when one describes an unknown object: "Golden Mountain". Here the object is not known culturally, however, the code already provides the necessary content and expression types. By combining the known semantic units one arrives a new expression (Eco, 1976, p. 187).

The rule-changing productions in contrast to the above require mapping into the expression from a content, which has not been analyzed into recognizable units yet. This is a paradoxical situation where the producer of the sign has a vague idea of what to say but does not know how to say it, on the other hand, s/he cannot know how to say what s/he wants to say until discovered precisely what to say. In many creative processes, such as artistic production, this is exactly the case. Only after an iterative process of trying out successive productions one would arrive a code which correlates the new invented expression with a new proposed content (ibid, pp. 188-189).

In IR one can come across both types of labour. The following sections try to identify the cases of *rule-governed* and *rule-changing* productive labour in the document retrieval situation.

[4] The concept of relevance is used provisionally at this stage. The act of making relevance judgement needs to be examined carefully. It will be shown in section 5.5.5 in discussing the rule-changing acts in documentary information retrieval in detail that, relevance judgements are not independent of more basic labor of *rule-following* and *rule-changing* (see section 5.5 below).

5.5.1 Rule-governed activity

One can identify rule-governed productions in IR at both the denotative and the connotative levels.

To successfully produce a sign-vehicle at the denotative level, the sender needs to know the rules which enables the correlation between the two semiotic planes; i.e. the rules of coding.

By a technique similar to *programmed stimulations* (cf. 5.3.3), one then can aim to elicit information. In this mode the producer of the query statement (expression) partially (partially, that's why its not overcoding!) foresees the result of the search with a given statement, nevertheless can not predict the precise outcome (see also footnote 13). The producer follows the rules of the code (index language+matching function) as much as s/he can foresee its effects. However, it is almost always impossible to predict its effects precisely and therefore the process involves learning/discovering the particulars of the code/coding.

At the connotative level, that is, at the instance when the query statement is replaced by retrieved documents and the user examines the individual documents that the system presents in response to the expression produced, the situation could include several different productive labour, sometimes all at the same time.

To analyze the complex situation at this level, let us assume that the user makes a judgement in response to seeing individual texts as to their usefulness for her/his purpose. This is a simplifying assumption, since as we will see in 5.5.5, the user does not need to judge the documents according to the criteria of usefulness or use value alone. However to make our analysis even simpler, lets assume that, the judgement pertaining to the usefulness of individual documents is a dichotomous one in the form of relevant/not-relevant.

One can envisage a situation of one of the two possible cases when the user is prompted for relevance judgement: it can be assumed either that the user is making relevance judgements in accordance to the rules set out by the code (index language + matching function; cf. 5.6.1.3) or, *proposing* a new code/index language.

In the first case, the user is positioned by the request made by the system for 'relevance feedback' as an addressee of a denotative statement. This is a case of *didactics*. The user being an addressee of a *transmitted* scientific knowledge is positioned on a unequal basis with the sender of the of the information. The sender is the expert in the field, whereas the addressee is the student, the subject of learning (see Lyotard, 1984, pp. 24-25). The addressee of the scientific statement is not in a position to judge its truth value in relation to the query statement that s/he submitted to the system, but expected to judge its truth value in relation to her/his overall enquiry or information need (see Robertson & Belkin, 1978a, for delimitation of the 'query' from the 'query statement').

The addressee posited by the request for relevance judgement in the position of the student who does not know what the sender (expert) knows and required to produce a true statement about the referent of her/his query: the documents. The query is often referred to as the underlying information need to indicate a conceptual or cognitive phenomenon, as distinct from the statement that expresses it which is referred to as the query statement or the search statement. This requirement for production of a true statement (extensional semantics: truth value of denotative statements; cf. 3.5.14) is the condition for learning or becoming learned (ibid).

The reason for distinguishing the query statement from the query itself is that, the user may not know how to specify the query accurately and/or completely. This is the case when the user is not expert in the domain and/or familiar with the system (its index language, matching function, command language, etc.) as discussed by various researchers and summarized by Belkin (1980, pp. 136-139).

It is assumed in this situation that, the main reason for retrieving unwanted documents is user's inability to specify what s/he wants to retrieve accurately. The problem is to find out what the user wants or needs. It is a case of 'non-perfect information game'[5]. It is a performance of rule-following on the part of the user. It requires the user (the student of the didactics) to improve her/his competence (skills) with respect to the knowledge of both the domain and the system, that is, to *reproduce* the existing knowledge. The problem is to establish the truth value of the querywith respect to the documents and vice versa.

There could be a totally different situation in which the sender of the query is an expert in the domain and therefore able to formulate her/his query statement accurately (and completely). In this case, the user (expert) is in a position to judge the truth value of a given document in relation to a given query statement. This is the assumption of many retrieval experiments with test collections and panels of experts (see for instance, discussion of this situation in Saracevic, 1975; 1971).

This is a rule-changing activity, a 'perfect information' game in that, the indexing language and retrieval rules are being tested with an aim to change it. To distinguish this from other rule-changing activities in IR (see 5.5.2 and 5.5.5), it will be referred to as 'code-making' (code making as coding the documents for retrieval). The code-making activity is moderated by a group of experts in the field and subject to their consensus. In this respect it is not different from the general *game of science* discussed in 4.3.2.2.

This sort of activity is similar to those productive modes examined under the heading of inventions in 5.3.4 in the sense that, one tries to posit a correlation between some descriptiveterm(s) and a document, between an expression and its supposed content. The success of this correlation is subject to normal procedure of acceptance by the experts of the field (cf. the case of the subject of didactics discussed above). Once a consensus is reached it becomes an overcoded sign function. This sort of rule-changing activity will be taken together with the other types of rule-changing activity in IR in the next section.

5.5.2 Rule-changing activity

The rule-changing activities are creative processes similar to those of inventions and paradigm shifts in science. They need to be clearly distinguished from the rule-following or rule-governed labour in IR which is about transmission and reproduction of existing knowledge.

The rule-following activity in IR is about transmission and reproduction of established denotative statements. It is therefore, about consumption of information/knowledge by the apprentice, the student, the trainee, while the rule-changing activity involves decision-making by the experts. The former is a game of non-perfect information in the sense used in Game Theory (Gibbons, 1992; Rapoport, 1960; Rasmusen, 1989, p.45). The latter is a game of perfect information. The

[5] See 5.5.2 for definition of 'perfect' and 'non-perfect' information games.

difference between them is that, in non-perfect or imperfect informational games, the moves of the opponents are not known at all instances of the game, whereas in perfect information games they are (Rapoport, 1960, p. 62).

In non-perfect information games, such as the ranked coordination game (Rasmusen, 1989, p.45), the advantage is on the side of the player who can obtain more information about the other player(s) (Lyotard, 1984, p. 51). Translating this to the IR process; when one is dealing with the transmission of information, the problem is to get more information about the user and her information need, by learning about the background, personal traits, problem area, etc. (the task of user modelling) and by making her/him to produce true statements about the documents and their relation to her/his information need (the task of relevance judgements and relevance feedback) so that the correct information (documents) are transmitted (the task of PRP[6]).

In perfect information games all the moves of the players are transparent at the all instances of the game (Gibbons, 1992, p. 55; Rapoport, 1960, p. 63) and winning the game does not depend on obtaining more information about the opponents' moves, but in arranging the data in a new way, i.e. imagining new moves, changing and prescribing the rules of the game, switching from one game to another, or even inventing new games (Lyotard, 1984, pp. 51-52). These are about fixing the rules and meta-rules of the game (cf. 3.9). In IR terms, all information (at least in principle) is available to the experts in a given field, (cf. 5.6.2.3). Thus, the expert is in position (unlike the student of the non-perfect information game) to determine the rules and meta-rules of the game.

We have already come across two such perfect information games in the context of documentary information retrieval situation: paradigm setting (deciding the meta-rules that govern the validity moves in a paradigm, cf. 4.3.2.2), and code-making (deciding the rules of retrieval, i.e. the index language, matching function, etc., cf. 5.5.1).

In paradigm setting (cf. 4.3.2.2), one is projecting (see fig. 4.3) from the dense continuum of undifferentiated human activities, named as the social text in 4.3.2.1 onto a set of rules which regulates (controls) the permissible interpretations of the social continuum, thus setting the meta-rules of the game. Meta-rules of a paradigm allow certain moves while repressing (excluding) the others.

The case of code-making was discussed in the preceding section, and it concerns with the controlling the vocabulary, therefore meaning (of the documents) in a given paradigm. In this case, one is mapping from the paradigm to the semantic model, thus setting the rules regarding the representation of the documents. As would be recalled from 4.3.2.3, the semantic model in IR consists of the documents plus the retrieval (representation) rules, namely, the indexing language and the matching function.

The first type of rule-changing in IR (i.e. paradigm setting) corresponds to the radical inventions (cf. 5.3.4) in other spheres of sign production activities. The second type (i.e. code-making) is similar to moderate inventions in that, being about the semantic model (content representation), it is less fundamental than the rules that govern the validity of moves in a paradigm (cf. 5.5.5). Another important type of rule-changing labour in IR involves knowledge production by establishing new connections between documents. The case of knowledge production will be treated in detail in 5.5.5.

[6] It will be shown in 6.1.2 and 6.1.3 that, the Probability Ranking Principle (PRP, see e.g. Robertson, 1977a) is exclusively concerned with this problem.

5.5.3 The political economy of knowledge consumption

This section is intended as a brief discussion of some of the social issues related to the retrieval of information from computerized databases. It is meant only to be an introduction to the complex socio-political space within which every information system exists as one among the multitude of codes that crisscross the social field. Any evaluation of information systems should take place within the context of this cultural space.

The analyses of the preceding two sections should make clear that the informational processes fall into one of the two categories: *consumption* and *invention*, or *transmission/reproduction* of the existing knowledge and *decision-making* regarding the rules and meta-rules of the scientific discourse. Therefore, the denotative statements about the truth value of a piece of information need to be methodologically differentiated from the prescriptive statements which set the conditions of validity of the former. Scientific knowledge does not constitute an indisputable truth that can be proven beyond any refutation. The rules of scientific knowledge are defined by meta-rules, which are prescriptive not denotative, therefore a result of consensus, rather than experimental verification. Science is a kind of language game (in the sense of Wittgenstein, cf. 3.9), its rules are fixed by its players (cf. 5.5.4).

The information game is part of this general game of science. There are decision-makers who are (at least in principle) have access to all the necessary information (hence, perfect information game; cf. 5.6.2.3) and concerned with various levels of rules that control the generation, dissemination and consumption of knowledge in the communication channels of the society (conceived and constructed as a system). The rest of the players of the game are sign-posted at the nodes of the social communication networks through which various messages pass. These are the consumers of knowledge, their role is restricted to the following of the rules, fixed by the decision makers (Lyotard, 1984, p. 15).

The consumers are not at the same level of competence with that of the rule-makers in that, they do not have access to perfect information. They play the game of non-perfect information which is concerned with producing denotative statements about objects. These statements can be declared true or false (admissibility of them is governed by the meta-rules prescribed by the decision-makers). The posts at the nodes of the network however, are not entirely powerless with regard to the messages that transverse them. They are allowed to respond to the messages creatively to a certain extent as long as it does not disturb the overall structure of the system. In a way, the system encourages the novelty of an unexpected move (in terms of language games), so that it can learn to regulate itself better by self-adjustment which helps to improve its performativity (ibid).

The society, viewed as functioning like a system by systems theorists such as Parsons (1967), is conceived as governed by the overall objective of optimising its efficiency, i.e. the principle of performativity. This is the positivist ideal of society as a giant machine, measurable in terms of the ratio between input and output variables of the system (ibid, p. 11).

The production and consumption of knowledge in societies run by the general objective of performativity is organized along the lines of other economic goods; those reserved for the decision-making in order to optimize the performance of the system, and those are for the consumption by the masses in their daily activity (struggle) of keeping up with the requirements of the system in their training and education. This is described as follows by Lyotard (1984, p. 6): "It is not hard to visualize learning circulating along the same lines as money, instead of for its "educational" value or political (administrative, diplomatic, military) importance; the pertinent distinction would no longer be between knowledge and ignorance, but rather, as is the case with money, between "payment knowledge" and "investment knowledge" -in other words, between units of knowledge exchanged in a daily maintenance framework (the reconstitution of the work force, "survival") versus funds of knowledge dedicated to optimizing the performance of a project.

If this were the case, communicational transparency would be similar to liberalism. Liberalism does not preclude an organization of the flow of the money in which some channels are used in decision making while others are only good for payment of debts. One could similarly imagine flows of knowledge travelling along identical channels of identical nature, some of which would be reserved for the "decision makers", while the others would be used to repay each person's dept with respect to the social bond". The social bond is linguistic.

5.5.4 Performativity or Paralogy

The positivist's principle of performativity which governs the rationale behind most of the scientific pragmatics has been showing signs of inadequacy roughly from about the middle of this century (see e.g. Feyerabend, 1978; also Wersig, 1993, pp. 234-235).

Scientific pragmatics (research and teaching) 'legitimize' itself at the final analysis, by 'performativity', according to Lyotard. The ability to produce proof which is one of the conditions of the scientific statement, increases by increased performativity, increased efficiency: "... since performativity increases the ability to produce proof, it also increases the ability to be right: the technical criterion, introduced on a massive scale into scientific knowledge, cannot fail to influence the truth criterion" (Lyotard, 1984, p. 46).

The problem of self-legitimation in science, that is the problem of justification of the prescriptive rules which determine the conditions of truth is resolved by the technical criterion of performativity. This enables collection of increased amount of data with greater efficiency about the referent of a denotative statement, thus, increasing the ability for 'context control', i.e. ability to be right. The question of "How do you prove the proof" is resolved, in other words, by the ability to reinforce reality, as Lyotard puts "... it has the means to become a reality, and that is all the proof it needs" (ibid, p.12).

However, as Lyotard remarks, there is a limit to one can collect information about the referent of a statement (game of non-perfect information). The positivist's idea of efficiency depends on achieving context control, i.e., existence of stable systems that can be represented by a continuous function, so that its output can be predicted from a given initial state. This deterministic ideal of science has been challenged by (relatively recent) developments in science itself. As research in quantum mechanics, and sub-atomic physics, fractals and non-continuous functions, and such, indisputably show, the simplistic assumptions of the rationalist's science are not maintainable in light of today's variety of scientific activities (pragmatics of science).

Science can no longer legitimize itself by producing more proof and controlling the context of experimentation. It works now at the limits of decidablity and control, between determinism and nondeterminism (i.e. paradoxes). Current change in science is, according to Lyotard, away from production of 'the known' to 'the unknown', which has nothing to do with maximized performance. This is accompanied by a move away from transmission of knowledge to production of knowledge, to invention of new moves and new (language) games (ibid, p. 53-60).

Legitimation in (the new) science(s), thus, take(s) place through a sort of 'paralogy': "Parälogy must be distinguished from innovation: the latter is under the command of the system, or at least used by it to improve its efficiency; the former is a move (the importance of which is often not recognized until later) played in the pragmatics of knowledge. The fact that it is in reality frequently, but not necessarily, the case that one is transformed into the other presents no difficulties for the hypothesis" (ibid, p. 61).

Lyotard, thus identifies two types of invention in the pragmatics of the *new sciences*: a new move, an innovation which is in the final analysis about extra performativity, and a new game or change in the rules of the game, which is incidental to the improved performativity. The important point here is that both of them are perfect information games, that is, about imagining new ways of arranging the data, and it is rather pointless to separate them except for methodological reasons.

5.5.5 New moves in IR

From the point of view of language games (cf. 3.9), communicational acts, i.e. utterance of verbal and non-verbal signs, are to be understood as participating in a game, similar to say, chess. Like chess, what Wittgenstein calls language games can be defined in terms of their rules-of-usage. Each language game is therefore, tells us how to use a particular word, a particular sign, to do things with it. Meaning of word is the use it is put in. When we know how to use a word, we know its meaning. Without a use, a word does not have meaning (Brown, 1974, p. 116).

There are various language games, the following are a small sample: asking a question, making a joke, making an hypothesis, making riddles, making an assertion, giving commands, reporting an event, guessing riddles, instructing, and so on (ibid, pp. 34-35). These are; interrogatives, performatives, prescriptives, evaluatives, denotatives, etc.

We have examined two of the above in some detail in the previous pages. A denotative utterance makes a true statement about its referent. A prescriptive utterance can be modulated as; orders, commands, instructions, recommendations, requests, pleas, prayers, and so on (Lyotard, 1984, p. 9).

Every utterance positions the sender, the addressee and the referent in a specific way. In denotations, the sender is the knower, the addressee is in a position to give or decline her/his assent, and the referent is positioned as something that must be correctly identified. In prescriptions the sender is the authority, and expects the addressee to perform the action referred to (ibid, p. 8).

There are practically infinite (indefinite) number of games, as even a slightest modification of a rule of a game, alters the nature of that game, resulting in a new game. Every utterance in a game is a move in the language game (ibid, p. 10).

The present section attempts to answer the following question: "what are the possible moves in the IR situation?".

One of the main and by far the most frequently performed game (speech acts) in IR is that of didactics[7]. This has been discussed in 5.5.1 (see also 5.6.2.3). Suffice to reiterate here that, the addressee is positioned as the student and required to produce a true statement about the referent of her/his information need. That is, the user is supposed to be able to establish whether or not it is true that a given document is referred to by the underlying information need, the problem need to be solved, the task that to be performed, here and now. 'Here and now' is crucial and it distinguishes it from the inventive moves. Here and now means that what is required is to have the competence to reproduce the existing meaning (interpretation) of the document, recognize only the established significance of the text in relation to the problem at hand. It reduces the indefinite ways of assigning meaning to the document to a controlled set of known and valid ones. The established meaning(s) of the document is presented to the user as an indisputable truth(s) and the user (the student) is asked to reproduce one of these meanings, here and now.

This is not the dialectic game of research between equal experts, but transmission of known facts to the addressee of the didactics, who is not at the level of competence to invent new meanings, interpret the text in a novel way, make new connections, or challenge the known meanings of the text. The student is not supposed to read it in a new way but to reproduce the known facts in relation to the problem that s/he needs to solve. In this way, the system not only posits the user as the producer of a denotative statement but reconstitute the problem in such a way that it can be assigned to one of the known problem types, i.e. can be solved by the existing knowledge.

In order the user to perform this task, s/he is supposed to learn about the domain of enquiry ("learn the facts") and the structure of the retrieval system. Some retrieval systems are designed to facilitate this explicitly (cf. 5.5.1).

It is not actually important whether or not the system itself positions the user in such a way by prompting to make a relevance judgement. It may be that the user positions herself as the addressee in didactics as a result of following more general set of rules, the rules set by the education system that constitutes the subjects as 'students' for instance.

There are however a number of other possible moves in the retrieval situation that does not constitute the user as the consumer of information/knowledge.

There are cases in IR, where the user does not seek to retrieve (transmit) an existing piece of knowledge but wants to seek new ways of interpreting them. The end goal is not the recognition of the significance (meaning) of a piece of text in relation to a practical problem at hand at present, but to make an innovative move whose significance may not be apparent until for some time.

In this case the extensional value of the relation between the text and the task one set to achieve, is not at stake, simply because it is a new move, a new proposition in the process of making whose admissibility is to be established at a later time -- if ever -- (as it is not a finished proposition yet). The value of a document is therefore can not be assumed to rest on being the referent of a denotative statement. Its value rests on its potential to be of significance at a later

[7] It can be argued that this is in fact the only sort of labour (with a few exceptions, such as O'Connor, 1993; Swanson, 1987; 1989; Davies, 1989; Bawden, 1986) that has been considered in the IR research so far. The difference between the rule-following and rule-changing pragmatics has not been explicated in IR literature.

time. Its practical value is deferred, indefinitely, until it is made pertinent for a new purpose.

One could envisage two possible such moves in a retrieval situation, which correspond roughly to moderate and radical inventions mentioned in 5.3.4[8].

A moderate invention, in the context of the search process, occurs when existing knowledge arranged in a new way, such that, eventually results (or aimed at resulting) in a change in the content model (semantic model). Such is the case in many research situation, where, immediate practical value of some retrieved documents may not be evident until at a later stage of the project. The project may result in publication of some research articles (therefore, production of new knowledge) and may well be put into some retrieval system, thus causing (infinitesimal) change in its content plane. Supposing that it deals with a new problem, yet staying within the established paradigm, it could be considered as a moderate invention.

A radical invention in IR search context is similar to the above. However, it affects not only the content model but the established paradigm as well. In this case, the implications of the move is radical in the sense that, it requires a change in the meta-rules which regulate the admissibility of every move within the domain (paradigm). Similarly, in this type of retrieval activity, relevance judgements with regard to the retrieved documents may not be immediately applicable, the value of the document to the proposed statement might not be established until the very late stages of the research process[9]. The value of any document to a radical new statement is not only a function of the 'internal time' of the move (i.e. time measured from beginning to the end of the move[10]), but also depends on the acceptability of the new proposed statement by the fellow experts in the field. It takes communication between the researches in the same paradigm until the proposed statement gains acceptance, that is of course, if it ever does. This means that, relevance of a given document to a radical statement cannot be predicted until the moves (arguments) of the other experts in the field are known, and may well change according to their responses. The relation between the newly proposed statement and the documents published in a field is, therefore, not fixed (stable) until the proposed statement finally gains acceptance. It may very well fail to gain acceptance at all, if it succeeds however, it is usually after much debate and even controversy, as it is the case with major paradigm shifts in the history of science (Kuhn, 1970; Feyerabend, 1978).

Since a radical move means a change in the paradigm or a change in some of the rules of the paradigm, it may be expected to result in organizing the existing research in the field according to a new pertinence/relevance criteria. This, as it is usually the case in the research situation, requires a dialectical debate among the experts, i.e. production of arguments and counter-arguments. Therefore, it may prove in the end that, the pertinence of a text is not what was foreseen initially, and requires re-evaluation. In its capacity of altering the pertinence of produced texts, radical inventions are not about reproduction of existing meanings (therefore,

[8] Here, inventions are considered in terms of the search process in IR. In section 5.5.2, inventions in relation to other aspects of the IR situation have been discussed.

[9] Hence, the relation between the query (statement) and the document is not that of a relevance or pertinence (or use value) in the sense of an *immediate* judgement, as practical value of a document is deferred (indefinitely).

[10] Any new statement is a result of an iterative process, where new proposition(s) are formulated and re-formulated, therefore involves various stages.

rules), but establishment of new prescriptive rules (meta-rules) that produce new linkages between the texts, in doing so, changing their relevance with respect to each other.

To sum up the above arguments, the criterion of usefulness or relevance of a document in relation to a given query in IR situation, is only *immediately* applicable in the case of denotative statements. The denotative statements are produced in IR specifically (although not exclusively) in the cases of transmission of knowledge considered as facts (from an established repertoire such as a database) to a non-expert (the student); i.e. in the didactics situation. In the case of didactics, the recipient of the knowledge is required to limit the problem within the established boundaries of the domain, therefore, this is a case of rule-following. The task of the user, posited as a student, is to establish a truth relation between her/his information need(s) and the texts s/he is presented with. This requires on the part of the user a double movement of fixing the problem and recognizing the significance of the documents in relation to it.

The relevance criterion is not however immediately applicable in the cases of prescription of new rules, whether as a result of moderate or radical invention. The user of an IRS who is involved with such an inventive labour, as in the many instances of research (knowledge production), is not bounded by the above mentioned double requirement of the addressee in the didactics situation. The researcher having a competence in the field of her/his research, is free to interpret texts in novel ways, establish/propose new links of relevance. In this respect, s/he is not limited to make denotative statements that describe/reproduce the relation between a fixed problem situation and its referents (the documents), but free to invent new ways to organize the problem, therefore the document(s) related to it. Inventive acts are characterized by 'thinking' which requires deferral of an arrived, fixed meaning, rather than instant transmission of information[11]. In such acts, one is not describing (denoting), but prescribing (setting new rules, standards, values, etc.). This is, in contrast to the rule-following activities, a rule-changing activity (see 5.6.2.3 for more discussion of inventive moves in IR).

Having made the distinction between the rule-following and rule-changing activities in IR, it is now necessary to note that, in practice it is more likely that, the rule-following and rule-changing activities are closely related and may not be so readily treated as two distinct modes of production. However, this demarcation is methodologically necessary (and potentially fruitful) in designing systems for document retrieval purposes.

[11] Thinking requires *deferral* because of one or some of following reasons: *a)* a document which is not useful itself may lead to another which is useful *b)* a document may not be useful on its own but may be useful when considered together with other(s) *c)* a document may not be useful itself as such, however, it could be necessary to know that document in order to be able to know the domain in general. Screening of such a document from the user may not be desirable (see 5.6.2.3 and 6.2.1.3 for more discussion of this point) *d)* until the moves of the other players in the game, so to speak, are known, it may well be impossible to foresee the value of a document. It might be necessary to backtrack and change ones' moves, changing the value of the documents in relation to the new position.

5.6 A communication model for IR

The preceding sections discussed the various aspects of document retrieval in some detail. This section intends to present the above arguments in a more coherent way, so that, the overall picture of IR situation would emerge.

The discussion in 5.5 challenged the simplistic model of IR where, information is assumed to be flowing from the user to the system and back in a *linear* fashion (fig. 3.2).

It has emerged from the ongoing analysis of this chapter that, communication in the IR situation is far from being linear and homogeneous (smooth), a *multilevelled discursive* (striated) process, involving different modes of production and different modes of interaction, sometimes all at once. Figure 5.5 summarizes this complex communicational process.

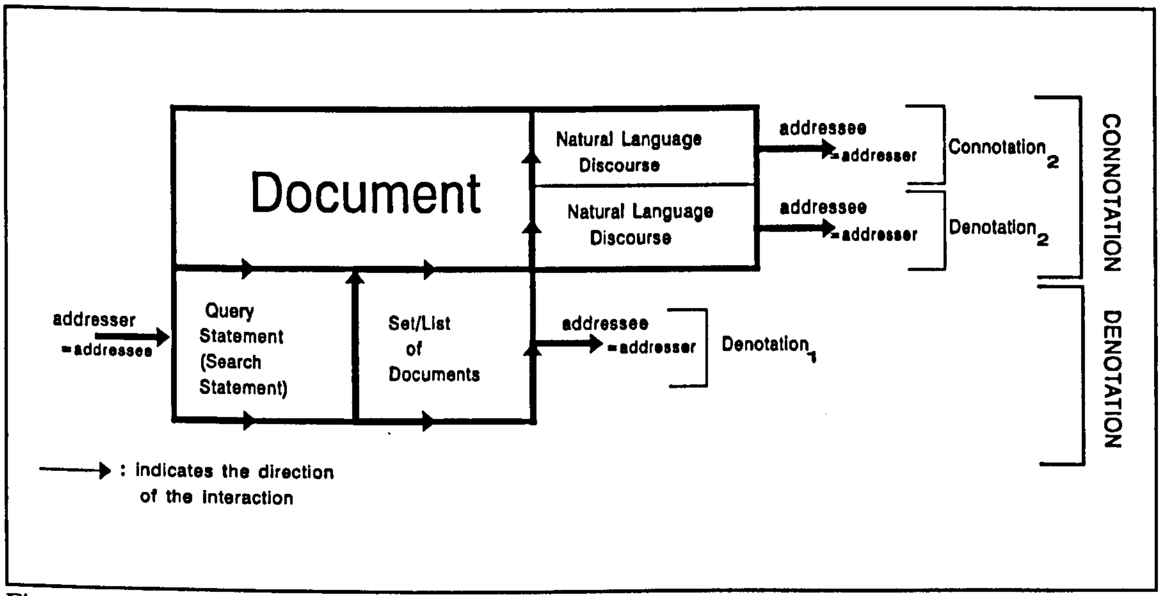


Figure 5.5: Communication in IR

5.6.1 Level$_1$: denotation$_1$

The first level of interaction in IR is comprised of defining the topological properties of the documents one wants to retrieve as discussed in 5.4.1. This requires from the user (the sender) a physical labour of production reminiscent of what has been called as programmed stimuli in 5.3.3.

5.6.1.1 The productive labour at denotation$_1$

Three important points that characterize this mode of production, should be noted: *i)* it is governed by a particular type/token ratio, called difficilis in 4.2.2.3, which accords an expression unit direct from a content-type *ii)* it is an undercoded sign-function in the sense that, the expression unit is correlated with a content which is not a well defined atomic unit, but a discourse *iii)* the sign producer, given that the content is not a unit but a discourse, can only predict part of the effects generated by the expression which function as its content (meaning).

This productive labour is not dissimilar to the so called iconic signs, in that, its production is governed by ratio difficilis and the content plane is not a simple unit but text corresponding to a discourse. Since the expression is generated according not to an expression-type as it is in natural language discourse, but from a content-type which is not analyzable into simpler units (cf. 4.5.3), it requires an inventive labour on the part of its producer. This means that, the sender of the expression needs to produce an expression to which no content-type is correlated. The producer of the sign-function therefore, needs both to invent the content-type and the expression corresponding to it simultaneously.

In IR context this means that, the user needs to re-construct semantic model and invent an expression which defines it. Once the semantic model is known, it is a matter of 'conventions of similitude' (cf. 4.3.1.2) to select the pertinent features of the 'semantic model' to produce the corresponding expression.

The semantic model is constructed from the underlying continuum of textual field which constitutes a paradigm. A paradigm, as discussed in 4.3.2.2, is an amalgamation of systems, artifacts, exemplar, etc., which are inseparably intertwined and continuously refer to each other. They embody a set of meta-rules, which prescribe the scope and content of the paradigm. The semantic model is constructed by abstracting the relevant features from this continuum.

In IR situation this means that, the activities constituting a paradigm are abstracted and represented in the form of published articles and stored in a retrieval system. The semantic model in IR therefore consists of a document collection and a set of rules that govern the representation and retrieval of the documents in the collection (i.e. an index language and a matching function, cf. 4.3.2.3). The producer of the sign-function needs to reconstruct this model and select the pertinent features to compose the expression. One therefore, in actual pragmatics of retrieval, starts from the text, and not from the query as an abstract mental construct. However, the user in most cases could only (re-)construct fragments of the text (i.e. the semantic model) to which the expression tokens are accorded. For this reason, the effects of the expression (i.e. its content) could only be foreseen partially (see footnote 13). The (re)construction of the text (semantic model) is an inventive labour, and the process of the formulation of the expression is an iterative procedure, similar to the creative productions described in 5.3.4 and 5.4.2. The production, however, also follows the rules of the correlational structure, namely, indexing and retrieval rules of the system, therefore, resembles to replicable expression functives of section 5.3.3, as well[12].

5.6.1.2 Correlation: equivalence or interpretation?

The user positioned at *level$_I$* produces an expression from her knowledge of the domain taking into account the indexing and retrieval rules of the system. This is a case of complex inferential process which requires hypothesis formulation and verification.

The user has to hypothesise about the semantic structure of the system (i.e. domain knowledge, index and retrieval rules) and from this hypothesis infer the result, i.e. the retrieved set (or ranked list). The hypothesis needs to be verified in light of the search result and changed if it does not conform with the result. The user interprets the expression in light of the observed

[12] Production of expression at this stage of IR is similar to *programmed stimuli* in this respect, and shares with it the peculiarity of lying somewhere between replicas and inventions (Fig. 5.1).

result. This is a type of inference Pierce called *abduction*: "In the case of hypothesis or abduction there is the inference of a case from a rule and a result" (Eco, 1976, p. 131). Abduction is more complex than simple inferential processes of deduction and induction (see fig. 5.6).

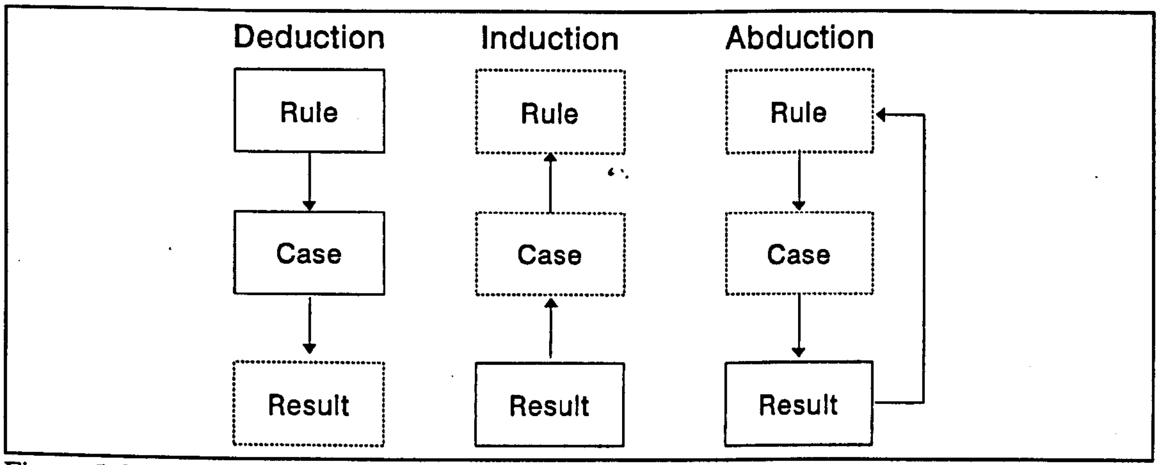


Figure 5.6: Deduction, Induction, Abduction (in Eco 1984, p. 40)

In deductions, one infers a result from a given rule and a case. In semiosis, deduction happens in cases where the correlation between the functives is that of simple equivalence, such is the case with the Morse code . In inductions, one infers a rule from a given result and a case. Such is the case when one infers meaning of a foreign word through repeated experience (Eco, 1984, p. 39).

In many acts of semiosis however, the rule to follow to make an inference either does not exist or not known. It has to be invented, i.e. posited or hypothesized and verified. When there are a number of candidate rules from which the most plausible one is to selected, we have what Eco (ibid, p. 42) calls an *undercoded abduction*. In undercoded abductions, one selects a rule according to contextual and circumstantial determinants. This has to be tested to verify.

In *creative abductions* however, the rule has to be invented and posited. This is the case with many interpretive labour in sign production, as well as, scientific inventions (ibid, pp. 42-43). "Many cases in which language is used not to confirm but to challenge a given world view or a scientific paradigm, and to decide that certain properties cannot belong any longer to the meaning of a given term ... require an interpretative cooperation that displays many characteristics of a creative abduction" (ibid, p. 43).

In IR situation, the user needs to refine his search in a trial and error procedure, making and changing hypothesis about the possible content of the expression. The user cannot foresee all the possible correlations of an expression (the result). There are two reasons for this: *i)* the correlational rules are usually complex, thus more than mere equivalences (see 5.6.1.3); *ii)* the content of an expression in IR is a discourse and not a simple unit. In short, the rules (indexing and retrieval) need to be abduced from the observed result.

5.6.1.3 Correlational structure as an interpretative system

The above discussion brings us to the consideration of the role that the system plays in the interpretative process in the production of query statements in IR.

It has been noted in the preceding section that, the user produces expression in IR according to the inferential rules of the system. In undercoded abductions, it is possible to abduce the meaning of a sign by virtue of contextual and circumstantial selections (cf. 4.4.1). In IR situation, although inferential rules are fixed and limited in number, the user cannot easily predict their result completely[13] for the simple fact that, the outcome is not a word or a morpheme, but a complex text. For this reason in IR, the productive labour is more like a creative- than an undercoded- abduction. Although the rules (in theory) known and limited in number, their effect cannot be predicted completely.

In this respect the role of the IRS is not strictly a correlational (equivalence) structure. It embodies series of inferential instructions or discursive rules (cf. 4.4.3). The inferential instructions are of two types: *A)* contextual-rules and *B)* discursive-policies[14].

A) Contextual rules govern the operations performed on the units of either the expression or the content plane. The units of the expression plane, as it would be recalled from sections 4.1.1 and 4.5.3, are the terms that constitute the query statement. Similarly the units of the content plane, i.e. documents (cf. 4.1.1), are composed of smaller units of linguistic items, such as words, morphemes, etc. (cf. 4.5.3). The contextual rules perform operations of selection, weighting, re-formulation and so on, on these units. The following is an incomplete list of some of the components or operations that perform contextual selections in IRS:

- *Indexing rules*: Indexing is the process of representing a document with a set of terms. Indexing operation involves selection of terms from the document to be indexed and/or a repertoire of controlled vocabulary and assigning the selected terms to the document as searchable (i.e. pertinent, invariant; cf. 4.1.1) units. In this capacity, indexing involves prescription of rules that determine the representation of the documents in the database. These rules determine the index language of the system. In assigning a term to a document, indexing rules take into account both the overall context of the database, i.e. subject(s) stored in the database, and the context in which the term appears in the individual document. I will refer below to the first of these contexts as 'inter-document' context and the latter as 'intra-document' context for convenience. In short, indexing rules perform selection operation on the units of the content plane.

[13] The user positioned at denotation$_1$ produces the query according to the form of the documents to be retrieved, therefore the results at this stage should be understood in identical terms, i.e. in terms of the *topological* properties of the documents. When it is said above that, the user cannot predict the result of her/his expression, it is therefore meant that, all combinational possibilities of the query terms cannot be predicted (alternative/complementary definition of the content of query terms at denotation$_1$ is given in 5.7.3, cf. also, footnote 43 in chapter 4). However, it should be made clear that, the results should not be understood in terms of the form of the retrieved documents alone (this is why in 5.6.1.1, it is said that, denotation$_1$ is an undercoded, rather than overcoded, sign-function, cf. 4.5.3). It is rarely, if at all that, the denotative level (denotation$_1$) in IR works in isolation with the next level of coding, called the connotative level. When the user produces an expression she often anticipates its results in terms of one or both of the codes of the next level as well (cf. 5.6.2). This should be born in mind in the rest of the discussion regarding sign productions in IRS.

[14] Note that, discursive-policies is one of the two constituent categories of discursive-rules (or inferential instructions), and should not be confused with it.

. *Auxiliary tools*, such as, thesauri, go-see-lists (GSLs), stopword lists, etc : These tools are used for various operations on the units of the either plane in connection with intra- and inter-document contexts. A thesaurus can be used for query (re-)formulation, as well as for indexing. Both of these operations involve contextual selections depending on intra- and inter- document contexts. Similarly GSLs, stop-lists, etc., operate on the units of the either plane, performing contextual selections that affect the outcome of the search.

. *Weighting functions*: By weighting function it is specifically meant here, statistical functions used for weighting search terms (Robertson & Sparck Jones, 1976; Robertson, 1981). These functions assign value to search terms which reflects the term's importance in inter- ('inverse document collection frequency' or 'idf') and intra- ('within document frequency' or 'wdf') document contexts.

All of the above operations and tools that perform contextual-selections are discursive in the sense that, the choice of a particular rule for contextual selection is determined by political decisions related to particular value judgements. Frohmann (1990) shows that, indexing rules are constructions (i.e. prescriptions in the sense of the present dissertation) which serve particular 'social practices'. Similarly, several weighting functions have been devised by making various explicit or implicit assumptions about the discrimination value of the terms and the relevance criterion. The relevance criterion is shown in section 5.5.5 and 5.6.2.3 to be highly political, in the sense that, it is inseparably intermingled with the overall policies of the design.

B) Discursive-policies are simply explicit or implicit criteria that determine the retrieval rules or the matching function. The simplest of the all matching functions are the Boolean and 'best-match' functions. However there are explicit models that relate the matching function to the effectiveness of the IRS. 'Utility-theoretic' models (Cooper, 1973) for instance, assign value to documents according to their expected utility and the matching function (sometimes, indexing rules, as well) are determined to reflect these values. These explicit models make specific assumptions about the function of the IRS and the relevance criterion. Such assumptions, as noted earlier, are fundamentally political (cf. 5.5.5 and 5.6.2.3), and in this respect linked to the 'political economy' of knowledge that was described schematically in 5.5.3.

It should also, be noted that the contextual rules discussed earlier are not always independent from the discursive policies (retrieval rules) that determine the outcome of the search. All of the operations in category *A* can be directly or indirectly related to the matching function employed for retrieval. The purpose of characterizing various operations performed in the IRS in terms of the above categories is to simplify the below discussion regarding the labour required on the part of the user in performing an IR act.

It is interesting in this connection to consider, whether it is the machine (the system) or the user who does the inference in the final analysis? The fact that, inferential instructions (inferential rules) can be programmed into computers does not necessarily implies that, machines can 'think' or 'interpret'. Even simple structures (such as, 'institutional codes', Eco, 1984, pp. 179-182), or written texts (Warner, 1991) may contain 'interpretive instructions' (inferential rules). It is a flawed argument to attribute intelligence to rules in that, without a query statement, without some sort of purpose or expectation, there is no code, or coding, i.e. there is no semiosis. It is the user who makes use of a coding system (IRS)[15]. The main hypothesis of the semiotic model developed in chapters 4 and 5 is that, it is the user who performs the inference to

[15] See Brier (1992, p. 102) for a similar argument: "... there is no information without mind, but there is no mind without nature, and there is no meaning without society and culture".

determine the meaning (content) of the expression. In doing so she has to take into account the rules of indexing and retrieval (matching function), i.e. contextual rules and the discursive-policies of the system. The user, for instance, needs to take into account the specificity of her search statement's terms, the linguistic contexts that they might appear in (the combinational possibilities of the terms), the index language of the system, as well as the matching function itself, to predict the outcome of the search. The inference labour is firmly located with the user, in this sense[16]. However, more complex the inferential rules are[17], less of a decipherable code an IRS becomes[18].

It is flawed to attribute reasoning capability to the system from the technical point of view as well. The inferential rules embodied in the system are fixed. It is fixed by its program (the software), which is a piece of text[19]. It is therefore determined by its programmer(s). Its rules can only be changed by re-programming. One can argue that various so called intelligent, adaptive systems, such as those based on neural nets change their programme by learning, modelling, etc. However, to this date there exists no system, which changes or overwrites the meta-rules which determine its working[20]. Therefore, the system is bounded by its program, by its programmer(s). For this reason, it is from the perspective of the semiotics adopted here rather pointless to ascribe intelligence to rules and artifacts that embody them.

A comparison of how rules change in natural and artificial semiotic systems should throw more light into this discussion. In natural semiotic systems, change occurs with usage. The users of a natural language invent new rules of usage, invent new games, new terms, etc (cf. 5.5.5). This causes incremental change in the language-system (*la langue*) over a period of time. Without usage there is no language-system (cf. 3.5.6).

In artificial semiotic systems, such as IRS, the change occur by changing the program (change in steps, in contrast to continuous change) (see Andersen, 1990). This could be a result of a single person's decision, or more usually, of a group of people's (cf. language games in 3.9 and 5.5.5). It is not unusual however that this is, at least to a degree, affected by the behaviour, response, etc. of the actual users of the system. In this respect, there is not much difference

[16] van Rijsbergen (1989) advocates a similar view, when saying: "I am offering a tool designed which I conjecture, when in the hands of a user, she will *learn to use* and she will *learn to do things with, ...*" [my *emphasis*] (p. 86).

[17] The general tendency is such that, especially the discursive policies, which increasingly determine the contextual rules (see e.g. Robertson, 1979) as well, have become more and more complex. This is especially conspicuous with the document-by-document (single-item oriented) view (Bookstein, 1989). See also van Rijsbergen (1989): "... the tool is so refined that the user cannot understand it, e.g. Probabilistic retrieval" (p. 85).

[18] See also 6.2.1.1, where, the inference labor required from the user is discussed in the context of IRS design problematic.

[19] See Warner (1991), for an argument that software is a sort of 'writing', a text not unlike documents, therefore attributing intelligence to a linguistic output is unfounded. See Andersen (1990) for a similar argument.

[20] This is not to imply that they could not possibly ever exist one day. However, see Dreyfus (1992) and Searle (1984, pp. 28-41) and Winograd & Flores (1986, pp. 99, 100-106) for strong arguments against the possibility of this.

between the two sorts of system. There is only system if there is usage. It is therefore, more appropriate to say that, not the system but its users, the actual usage embodies the inferential rules, and not the artefact. The programmer(s) can invite certain interpretations, encourage certain uses, but it ultimately depends on the users to attach meaning to the expression-vehicles that are manifested by the computer on its screen (ibid).

5.6.2 Level$_2$: connotation

Once the system replaces the query terms with some documents (metonymical operation on the part of the system), the user switches from one code to another, from an artificial coding, to the natural language code. This level relies on the previous one in the sense that, it is temporally after it. However, they are much closely intermingled than this. There are two major discursive operations at this level, corresponding to two sorts of cause-effect relation between the two levels, and can be categorized as strong and weak relations. The stronger relationship is called denotation$_2$ in the communication model of section 5.6 . The weaker is called connotation$_2$ (fig. 5.5).

5.6.2.1 The productive labour at denotation$_2$

At this mode of interaction the user is positioned as the receiver of transmitted knowledge, the addressee in didactics. The user is required to (learn to) produce correct denotative statements about the referent of her information need, the referent being the document in question (cf. 5.5.5). In doing so, she recognizes the document as being the referent of the 'query' (the 'Q-I-D model' of chapter 4).

The following features characterize this mode of production: *i)* the meaning potential of the document is reduced to a single (or a few) meaning(s) that are culturally known *ii)* the user recognizes the document, which now has a fixed meaning, as an artefact (object) to which the query refers *iii)* the document may be viewed by the user as an expression of the class of which it is a member in a process similar to ostensions.

There are several operational IRS that employ a variety of techniques to facilitate interaction as described above. It is worthwhile to discuss here two particular techniques used in IR that explicitly or implicitly assume the above described productive labour (i.e. recognition and ostention) on the part of the user.

In systems which employ document clustering, categorizing and classification (Deogun & Raghavan, 1986; Salton & Wong, 1978; van Rijsbergen & Croft, 1975; Goffman, 1969), the documents judged by the user as being relevant or useful for his purpose are registered by the system to be a member of a class of documents which share some common properties. The class is defined by some 'similarity measure' (Salton & McGill, 1983, pp. 124, 216 etc.). In so far as the selected document by the user is an example or sample of the class of documents that it is a member of, this productive labour is similar to ostensive operations (cf. 5.3.2). The type/token ratio is ratio facilis when the expression (i.e. the picked document) is viewed as belonging to one of the classes of expressions, yet subject to ratio difficilis when it is taken as referring to the original document of which it is the surrogate, or its double (cf. 4.2.2.2 for discussion of doubles). This, as already mentioned in 5.3.2, is the characteristic of the ostensive labour.

Variety of systems however, employ some sort of 'relevance feedback' mechanism, where the user is required to recognize the document not in relation to a class of documents, but as an expression of a given content or meaning. This is a productive labour similar to recognition of symptoms and clues (cf. 5.3.1), in so far as the documents are viewed as objects (therefore, closed systems with fixed meaning(s), cf. the counter argument for that in 4.2.3) that are already produced and exists among other objects as expressions of pre-existing and coded relations. The documents, being natural language texts are correlated to their content (which they are the expression of) by ratio facilis.

However, although some similarity exists between.recognition of documents and symptoms, and clues, this is clearly a different kind of productive labour than those subsumed under the general term of recognition. The most important difference being that, (identity of) the agent (i.e. the author) responsible for the production of a document has not direct relation to the meaning of the expression. Where as in symptoms and clues, it has a much closer and significant association. Whatever the differences may be, the important fact is that, the document in this mode of interaction is viewed as a ready made expression of a coded content, which is the pertinent characteristic of the labour involved in recognition of symptoms and clues in differentiating them from other productive labour discussed in section 5.3.

In the didactics situation, the user being a student is not at the competence level necessary to recognize completely the information needed to solve a problem or to perform a task (see Belkin, 1980), therefore, IR systems are designed to facilitate this process (cf. 5.5.5). The following section examines the techniques that are used to model the user in the context of the productive labour at *denotation*$_2$.

5.6.2.2 Modelling the addressee in didactics

Modelling the user in IR takes different forms. In the context of this section, modelling is used in a broader sense, not only incorporating those systems that model user characteristics, search behaviour etc. (cf. 1.2.1.4), but also, any system that elicits some information from the user to retrieve documents.

One such technique that elicit information from to the user to resolve the information need has been proposed by Belkin and a system based on its assumptions is (partially) implemented (Belkin et al., 1982a; 1982b).

As noted in the beginning of chapter 4, the main argument of this approach is that; "there is an underlying informational need beneath every[21] query, and this can be expressed in a some sort of formalized language". Although, it emphasize iterative interaction with the system, it still advocates some form of formalized representation of knowledge (of the user's need and the documents), which is assumed to represent the problem more effectively. There are a number of problems with this approach; I will only pinpoint those more relevant to the present

[21] Every intentional enquiry. The ASK ('Anomalous States of Knowledge') hypothesis (Belkin et al., 1982a; 1982b) does not mention browsing activity, and it is not clear to me whether it considers this as a purposeful activity. Regardless of this, it clearly commits itself *exclusively* to goal-oriented activities.

discussion[22].

From the perspective of the present discussion the idea that information need can be represented more efficiently by another set of symbols other than the actual process of retrieval is highly problematic.

The language games perspective clearly argues that, the use of a language is its meaning, and this can not be reduced to propositional or definitional conditions. This is the view of Wittgenstein[23] on the relation between language and logic: "In Wittgenstein's view, propositional conditions are not given with rules-of-usage, for they cannot be shared-they cannot be conventional" (Brown, 1974, p. 27). Thus concludes Brown, rule-of-usage is unrelated to 'propositional rule-of-usage' (ibid). Therefore, one can conclude that (cf. 3.9), the logical-positivists' idea that an act of language usage can be adequately represented by an artificial language external to the language use at question is unwarranted: "His (Wittgenstein's*) understanding is clear: any language, be it artificial or natural, is understood not in terms of some other language, but in terms of itself ..." [*my remark] (ibid, p. 17).

This is the very same reason that Andersen adopts a semiotic approach to design of computer systems: "... the principle (i.e. the principle of immanence of 'structural linguistics'*) demands that a language should be described as a structure *sui generis*, and not a projection of something else, be it psychology or logic" [*my remark] (1990, p. 7).

The argument against logical-positivists' perfect languages can be substantiated from the structural semiotics perspective as well.

As it has been noted in 3.5.6, according to the structural linguistics' point of view, there are only *differences* in language, i.e. signs stand out against society's ability to equate them with each other (Eco, 1976, p. 72). In other words, people use signs to refer to other signs, as if they can be substituted for each other. Yet, there is always difference in language, the semiotic process, semiosis, never comes to an absolute termination (cf. 3.5.2). It has been shown in 4.5.2 that, IR as a code is structured in two layers (denotative and connotative), and the relationship between the two codes is very complicated due to the fact that, the denotative code (denotation$_1$), being governed by ratio difficilis is totally alien to the natural language discourse (governed by ratio facilis) of the connotative code. The gap (difference) between these two codes alone makes possible IRS to function as a signifying structure (cf. 4.5.3). IRS work as a communication medium not because we can equate the query terms with documents, it works because they are different. This makes perfectly clear that, replacing the actual act of interacting with the system with another act, that is, some sort of mediation through yet another set of signs and code (intermediate coding), is no match for the original code and the process of language use[24] that goes with it. Intermediate representation is not the equivalent of using the system to retrieve

[22] For a thorough treatment of the problematic approach of the so called 'cognitive paradigm' in IR, the user should refer to Frohmann (1992). See also Winograd & Flores (1986, esp. pp. 23-26) and Brier (1992).

[23] The logical positivists' idea of 'perfect language' is rejected, apart from Wittgenstein, by Austin (1962), Searle (1979), and others.

[24] From the point of view of the semiotics adopted in this dissertation, IR systems function as a code, a *language* on its own right, and the interaction with the system for retrieval purpose is effectively corresponds to *language usage*.

documents. At best, it constitutes another code, which is not the substitute for the original.

Therefore, the cognitive viewpoint's overemphasis of modelling of the user's need is hard to justify. There is no substitute for the actual use of language after all. However, it is perfectly okay, to some extent (cf. discursive policies in 5.6.2.3), to try to help the user with the system's structure and the structure of the domain of enquiry in conjunction with the didactics situation. This leads us to the second problem area with the cognitive viewpoint, that it is narrowly focusing on the cases of didactics (the cognitive viewpoint is not alone in this respect; the so called system oriented paradigm (Ellis, 1992), including the relevance feedback and the PRP approaches take a similar orientation, cf. 6.1.2).

The second problem with the cognitive paradigm is its inherent underlying assumption that IR is[25] concerned exclusively with transmission of information. To quote from Belkin "... IR, is a problem-oriented discipline, concerned with the problem of effective and efficient transfer of desired information ..." (Belkin, 1980, p. 133). The implicit assumption in this quotation is that, there is a problem that needs a solution now, and to solve the problem all is needed transmission of some information that is missing from the user's state of knowledge. Belkin states explicitly later on in the same text: "The success of the communication is dependent upon the extent to which the anomaly can be appropriately resolved on the basis of the information provided" (ibid, p. 135). This is, no doubt, description of a situation exclusively concerned with didactics.

The problem with IR concerning exclusively with the case of didactics is that, it aligns itself with the ideology of commodification of information (cf. sections 5.5.3 and 5.5.4) without any debate or critical reflection. It also causes total erosion of the distinction between decision-making and rule-following or transfer of information, therefore makes impossible to discuss the needs of those who not only want to be informed, but to be able to make informed[26] decisions (see 5.6.2.3).

5.6.2.3 The productive labour at connotation$_2$

This mode of interaction involves inventive labour on the part of the user. The user being the expert seeks to read documents in new ways. Reading is used here to denote that it is an interpretative labour which involves the process of negotiation with the text, and therefore, some sort of space, temporal distance, and hence deferral of an arrived, closed, meaning is needed. It implies an unfinished process. Fiske (1982, pp. 3-4) describes 'reading' as: "... the process of discovering meanings that occurs when the reader interacts or negotiates with the text. This negotiation takes place as the reader brings aspects of his cultural experience to bear upon the codes and signs which make up the text".

In this respect it is radically different from the recognition of texts which was called as denotation$_2$ in 5.6.2.1 . One should note that at denotation$_2$ the text is a ready-made message

[25] Note the stress on 'is'. The cognitive paradigm (and the system oriented paradigm for that matter) does not *explicate* the difference between information consumption/reproduction/transmission on the one hand, and information/knowledge production/decision-making on the other.

[26] Knowledgeable.

which is transmitted and reproduced, in this regard, similar to iconic signs and images[27]. Reading on the other hand is a creative labour and involves to varying degrees re-writing of the text. From this perspective it is the opposite of (passive) consumption/transmission of images. It requires hermeneutic process, i.e. interpretation "... that sets that document in the horizon[28] of the subject knowledge and its potential application for a particular knowledge-seeker, who operates in an horizon of beliefs and practices" (Froehlich, 1989b, p. 63). The meaning of the text in this hermeneutic process is determined by social practices in an historical context, as Brier (1992) points out, referring to the work of Winograd and Flores (1986).

The following characterizes the labour at this mode: *i)* there is no immediate extensional value (truth-value, therefore use-value) of the document one is reading with respect to the query or the query statement[29] *ii)* it involves any of the following inventive labour which results in change either in the semantic model (cf. 4.3.2.3) or the paradigm (cf. 4.3.2.2): *a)* changing the indexing rules or the relations between documents (cf. 5.5.2; 5.5.5) *b)* changing the meta-rules of the paradigm (cf. 5.5.2; 4.3.2.2).

The idea of relevance judgement at a fixed point in time, which characterizes so many of the evaluation exercises in IR, is simply not valid in this mode (cf. 5.5.5). IR research and theory have so far missed the point[30] that there is an enormous difference between decision-making and reproduction of existing knowledge, between rule-changing and rule-governed activities. By a sort of metonymical transposition, the idea of fixed relevance judgements, which has a well defined role in context-controlled conditions of laboratory testing (cf. 5.5.1), that dominated the early IR evaluation tests (cf. e.g. The Cranfield tests in Cleverdon, 1967; see also Sparck-Jones in Sparck-Jones, 1981, pp. 213-255 for a thorough review of early IR experiments) has been transferred to the interactive end-user retrieval situation. As suggested in 5.5.5, in practice, it is hardly possible to separate the cases of decision-making and transmission of known meanings from each other. This is true even for cases of ordinary querying of databases, and certainly true for most if not all retrieval processes in the research pragmatics.

To illustrate the case of decision-making in the context of IR imagine the following situation[31]: A wine retailer wants to find the names of all wine exporters in France with a view to import

[27] There are certain structural similarities between retrieval of ready-made information in the situation of didactics and the transmission of ready-made *images*, say on television. Being governed, in their production and consumption by ratio difficilis is the most important one.

[28] Husserl's concept that means roughly: purpose, understanding. "The horizon is the unthemathized unconscious context for every single person that is the corollary to all his perception and consciousness ... Horizons are always restricted but they can move" (Hoel, 1992, in Vakkari & Cronin, 1992, p. 78), i.e. they change (cf. 5.5.5).

[29] Meaning that, use-value of a document is not *necessarily* determined at the time of reading (i.e. 'here and now', cf. 5.5.5).

[30] In this regard it is interesting to note van Rijsbergen's (1989, p. 86) comment on his recent proposal for a logic based inference model for IR: "This tool may not exactly correspond to a tool designed to *retrieve relevant documents* ... Thus it attempts to transcend tools based on *empirical* notions of *relevance*" [my *emphasis*].

[31] Partly inspired by an example in Bar-Hillel (1964, pp. 362-363). However, both the form of the example and the purpose it serves to illustrate here is completely novel.

from them. This underlying 'information need' prompts him to interrogate a database which has only two documents related to 'wine and France'. First document contains all the information required, i.e., the names of all wine retailers in France who export. The other document contains names of wine retailers in France who import foreign wine. Suppose now that our entrepreneur enters the following query statement to the system: "Wine France". Assume that, our system being very 'intelligent', distinguishes between the exporters and importers of wine, and ask the user ('understanding' or 'knowing' the meaning of both documents) if he is interested in exporters of wine in France. Our user naturally answers "yes" to this prompt. On the basis of the user's answer, the system concludes that the first document is the right (or most probably the right) one, and returns it to our user. The user looks at the document it obviously contains the right information, i.e. the complete list of wine exporters in France. This document resolves his information need, he thus leaves the system completely satisfied with the result.

Lets now suppose that, he is the director of a retail business, perhaps a one man business, and hence in charge of strategic planning of the whole enterprise, that is, responsible for the company's business plan. We can imagine that our entrepreneur has rather naive ideas about France (perhaps we can say that his image structure about France is ill formed or incomplete), and he thinks that France being such a big wine producer does not import any wine from other countries as the wine it produces is that of the highest quality and there is surplus of production. This misinformation could cost him dearly, for, if he only knew that, French companies do buy from foreign producers, being in a position of decision-making, he might well save his company from an immanent bankruptcy by enterprising to sell the cheap costing home-produced wine to the French importers. Our system being too intelligent, unfortunately screens[32] this information from our decision-maker, thus costing more than it saves for him.

This example might be an over simplified picture of the real situation (more realistic ones can always be constructed), however it does serve to illustrate the complex and sometimes conflicting issues of decision-making and optimizing the system's performance locally[33]. The major problem with many retrieval systems is that, they are designed according to decision-theoretic models that optimize their performance on the basis of a *linear sequence* of past events, thus producing *local optimization*, which could result in *sub-optimal* performance globally[34].

Compare the above example of the wine merchant with the case of a student, who refers to the

[32] See footnote 11.

[33] The argument that, it is always a matter of presenting to the enquirer a list of documents instead of another (Robertson & Belkin, 1978b) may not always hold true. The human intermediary, for instance, is much more flexible compared to a computerized system in terms of the number *language games* or *speech acts* that she can perform. To be advised and referred to some documents by an intermediary in a library is not identical to be presented with a *list* of documents by a computerized system. In the case of the human librarian, there is always the labor of *negotiating* and *discussing*, which is hardly the same thing as transmitted with a list/set of documents by the system. The latter is a frame-by-frame transmission of information (this tendency is especially strong in the document-by-document view, see 6.1.2), the former by the virtue of the process of negotiation and discussion (thus, thinking), opens up necessarily to other texts, and therefore can not be considered simply as being equivalent to a list.

[34] As the case of the wine merchant demonstrates, local optimization could produce detrimental results in the situations where decision-making labor has a priority over *efficient/effective* transmission of information.

same system with exactly the same information need and hence the query statement. Lets suppose that our student is given the task of finding names of all exporters of wine in France, as a part of a course work. The student submits the query statement "Wine France", goes through exactly the same process as the previous case of the entrepreneur, and similarly leaves the system totally satisfied. In this case, to a stark contrast with the previous situation, the student does not loose out any information as a result of the system's discursive policies (cf. 5.6.1.3). This is a case of transmission (reproduction) of information in opposition to the former case's production. It is similarly, an over-simplified situation, for, especially in research pragmatics, tasks are not as extremely about reproduction as this.

The above discussion of the two fictional cases makes clear that, the decisions to base the design criteria are fundamentally political. Design issues cannot be simply assumed to rest on technical criteria such as, efficiency or performativity. They are inextricably bounded with policy making. The policies of the above fictional system clearly favours transmission and consumption of information on the expense of the production/decision-making labour. It is worth noting here that, they are not trivially conflicting pragmatics. One cannot simply assume that it is a matter of collecting more information about the user, so that her decision-making options are *pre-known*, and the system's performance is optimized accordingly, i.e. the retrieval process becomes a perfect-information game with optimal solutions (cf. 5.5.1 and 5.5.2). Apart from this being either not implementable or economically unfeasible in practice, overwhelming majority of real-life decision problems are 'nonzero-sum games' where utilities of outcomes may not be determinable from the available information, therefore, the best strategy cannot be known (Rapoport, 1966, pp. 201-204)[35]. Even if it is known, as it is the case in 'zero-sum games', it can be known by any competitor, therefore, no advantage could be gained ultimately (ibid, p. 204). Advantage can only be gained in this case, by making a novel move; switching to another game, inventing a new game, so on (cf. 5.5.2).

In this respect, decision-making requires *redundant* information, that is, information required to make a decision can not be predicted *a priori*[36]. For this reason, efficiency and decision making are contradictory conditions. In the cases of production of knowledge (and decision-making), information does not necessarily lead to reduction in uncertainty (with regard to the optimal strategies). On the contrary, it can be argued that, in such cases, provision of information could lead to increased uncertainty (with regard to the utilities assigned to outcomes, therefore, with regard to the alternative strategies): "In fact, it is reasonably clear from research that in particular circumstances the provision of information could well lead to increased uncertainty" (Halloran, 1983, p. 160). As the example of the wine merchant demonstrates, new information could lead to increased levels of complexity, more possibilities, thus, more uncertainty, which is only resolved by the act of making a decision (a new move, a new statement).

The problematic of accommodating this sort of inventive (rule-changing) labour, with seemingly denotative (in general) problem of retrieval is addressed in the next chapter.

[35] See Winograd & Flores (1986, pp. 20-23, 98, 144-150) for a very similar argument to this.

[36] "The essence of intelligence is to act appropriately when there is no simple *pre-determination* of the problem or the space of states in which to search for a solution. Rational search within a space is not possible until the space itself has been created, and is useful only to the extent that the formal structure corresponds effectively to the situation" (Winograd & Flores, 1986, p. 98). Compare this with section 5.6.2.2, and especially with section 5.5.5, in relation to the double constraint imposed on the subject of didactics.

To illustrate the non-denotative nature of rule-changing labour consider the following situation: the documents that are to be stored in a retrieval system are first shown to an expert in the domain who reads the documents and indexes them. Being an expert, she understands their content and knows how they relate to the other documents in the same subject field that are already stored in the system (the case of perfect information game, cf. 5.5.2). For this user, searching the database should be a totally useless activity. As she knows each and every document in the field, she would not need to interrogate the database, except to retrieve a known item. In this case, since she knows all the indexing rules of the system, this is a trivial matter of course. Our expert on the other hand, most probably would endeavour other sorts of activities than searching (e.g. research) which may involve producing statements that inadvertently or not, influence the truth-values of the documents in the database with respect to the queries (cf. inventions in IR in 5.5.5). The overall activities of the expert in the above example can be viewed as an extreme case of the retrieval activity[37] which does not involve any denotative labour. The function of the user in the above described situation is to scan and make new connections[38] between the documents. This may well be an extreme situation, but nevertheless many research activity resemble to it in varying degrees, at least at some instances.

More realistically, the experts follow the literature closely and use the retrieval system occasionally to keep up-to-date with the new publications in the field. To 'keep up' implies a reading situation where the reader brings about her previous readings to interpret the text, which consequently results in a new arrangement in all past readings. This process has been referred as intertextuality in 4.2.3. For this user, it is clear that, the idea of relevance judgements as an act fixed in time and space would not make much sense[39]. All publications in her domain of expertise are relevant to this user[40] (perfect information, cf. 5.5.2). The expert is in the position to read the documents (ideally all in her field of expertise), to understand them, to interpret and re-interpret and connect with the other documents[41]. All these activities take place over a long period of time. Unlike the student of the didactics situation, she has to know the document and

[37] If not a retrieval activity in the sense of 'retrieving' documents as such, it is certainly a 'retrieval system design' activity.

[38] Indexing is about making connections between documents. The act of producing a statement is another such labor as, any statement refers to and re-arranges the relations between other statements, and thus the relations between the documents that contain them.

[39] Supposing that, the expert is familiar with the structure of the retrieval system (index language, matching function, etc.) as well. An additional condition for this to hold is that, the system (rather the index language of the system) should be capable of discriminating between any retrievable set of documents (Bookstein, 1989, p. 469).

[40] One may from this argument rush to the conclusion that, for this particular situation, the main problematic of retrieval is to define 'objectively' the boundary conditions of the subject domain. As this is fundamentally a consensual decision (i.e. cultural/social, not private) it is not possible to determine the limits of a domain in practice. More importantly, since this is one of the main activities in science (see below paragraph), it is rather hopeless to try to fix the boundaries for any practical purpose. Consequently, design criteria for IRS should/could not be based on this. In chapter 6, the problem of catering for the inventive labour in systems design in IR is discussed.

[41] Which is radically different from recognition of texts as describing (i.e. denoting) certain informational need. The latter is what the subject of *didactics* does.

assign meaning to it. The student does not need to, nor possibly could, understand or assign meaning to all documents in a given domain. Only a few, which are schematically about his information need, i.e. about the material he is assigned to learn about, could make sense to him (i.e. could be attributed with a known/established meaning).

As a conclusion to the above argument, it can be said that, the relevance criterion understood as a fixed *transcendent* judgement is not pertinent to the expert[42]. The expert scans or reads and make connections. The idea of relevance enters to the picture only as a process over time, and in relation to the boundary, therefore, the rules of the research paradigm (cf. 4.3.2.2, 5.5.2): "Where does my research interest stops, which documents are/should left (best) out of this boundary?". This is the problematic that concerns the expert. As discussed in 4.3.2.2, it is a consensual decision. Since it is a consensual decision, it is subject to change (sometimes abruptly), and consists the main 'paralogical' activity in science today (cf. 5.6.2.4).

One needs to explicate the difference between those activities resembling to the above described case of rule-changing and those involving didactics (learning) in order to appreciate the complexity of the documentary informational retrieval situation. The consequences of this demarcation in terms of systems design practice is discussed in chapter 6.

5.6.2.4 Research as a productive labour

This section describes the research situation as a productive process as opposed the reproductive process of learning, in the context of information retrieval.

As argued in the preceding paragraphs, the relevance judgement mechanism, as well as the more general mechanism of modelling is a discursive structure that configures the user in specific ways[43]. However, as it has been discussed in 5.5.5 and 5.6.2.3 that, there are instances of interacting with IRS, where the object of interaction is not the transfer of information as such, but the *creation* of new linkages of relevancies, i.e. organing the existing data in new ways.

There is hardly any explicit attempt in the entire history of IR that tackled the problem of designing a system that enables the user to help look at the retrieval process from, not the point of view of transmission/reproduction, but production (with a few exceptions, such as, Swanson, 1987; 1989; Davies, 1989; Bawden, 1986). The overall objective of this dissertation is to attempt

[42] All this is said, the problem of contextual ambiguity (ambiguity of terms as a result of their varying contextual use) could cause problems to our expert who is both familiar with the domain and the system, thus the 'relevance concept' in the narrow traditional sense could still hold a marginal role to play. Another important factor that could cause difficulty in retrieval is the possible existence of confusing and/or conflicting rules of indexing that might prevent the expert in our example from specifying exactly what she wants.

[43] Interaction should be understood in this sense. It a is communicational act (speech act, in the sense of Wittgenstein, Austin and Searle) that positions (configure) the user as one of its posts (addressee, sender, or referent) according to the particular language game that is programmed to perform. Since it is pre-programmed for a particular game (or a menu of games) it is not flexible in comparison to human to human language games, where the participants can (and often do) switch from one game to the other at each utterance. More importantly, the programme cannot invent or propose new games other than determined by its text (meta-rules), which are fixed, of course, at the time of programming.

to design an IRS (discursive mechanism) from the perspective of knowledge production (within the constraints set by the design environment, such as the available tools, resources, etc). This is fully discussed in the next chapter. It is the objective of this section however, to discuss the research activity, which is viewed primarily as a productive labour in relation to IR located in a broader social context.

Lyotard[44] in 'The postmodern condition: A report on knowledge' argues that logical positivists' idea of performativity (a technical criterion in itself) which is used to legitimize the scientific pragmatics has started from roughly about the mid twentieth century to show signs of inadequacy as a legitimizing agent with regard to developments in some areas of the scientific activity (cf. 5.5.4). The most important characteristic feature of this development, which Lyotard calls as the 'postmodern scientific knowledge'[45], is that "... the discourse on the rules that validate it is (explicitly) immanent on it. ...this inclusion is not a simple operation, but gives rise to "paradoxes" that are taken extremely seriously and to "limitations" on the scope of knowledge that are in fact changes in its nature" (Lyotard, 1984, pp. 54-55). According to Lyotard, this new mode of *legitimation* has as its basis "*difference* understood as *paralogy*" [my *emphasis*] (ibid, p. 60). He goes on to argue that "To the extent that science is differential, its pragmatics provides the antimodel of a stable system. ... Science is a model of an "open system," in which a statement becomes relevant if it "generates ideas," that is, if it generates other statements and other game rules" (ibid, p. 64).

Beardon (1994) taking 'AI' as an example, argues similarly that, there are sciences that entangled with the logical positivistic categories and boundaries, and there is a postmodernistic tendency in some others, which take anti-positivist directions. The initial stage in AI's development is marked by the efforts to mimic general human intelligence and intelligent behaviour. AI research described by its pioneers as "'experiments' and as 'first trials of previously untested ideas' ... The gradual growth of the discipline marked a new type of scientific endeavour with a new role for theory and models... AI became a kind of 'practical philosophy' where theories could be embodied in models and behaviour of the model was a new reality that could be interpreted in a way that was inconceivable before" (ibid, p. 9).

This epistemological break with positivism did not confronted without resistance from the scientific community: "The emerging discipline caused considerable disquiet in academic and research establishments" (ibid p. 10). In Britain, the Lighthill Report commissioned by the (then) Science Research Council declared AI "... as a rogue discipline that had no claim for autonomy" (ibid). This was partly the reason, according to Beardon, for AI to abandon its early anti-positivistic experimental[46] paradigm for a weaker version, which studies goal-oriented computing techniques to solve practical problems. The new turn taken by AI "... can now be seen as an attempt to map the territory initially explored in the name of AI in a positivist manner" (ibid). Beardon argues that, however, this re-mapping of the original AI territory with positivistic terms, did not go without any challenge, and another anti-positivistic tendency emerged within the paradigm came to known as Virtual Reality.

[44] Whose work on the condition of the scientific knowledge in the contemporary socio-political setting remains an important and frequently cited reference among those dealing the implications of technology in a wider social context.

[45] Wersig (1993, p. 234) similarly calls this new stage of science as *postmodern science.*

[46] In the sense of the *avand-garde* in art.

This sort of legitimation by *paralogy* in science is not an isolated condition. It is the state of affairs taking shape in general in the social sphere (i.e. socio-political organization of the society at large): "...the temporary contract[47] is in practice supplanting permanent institutions in the professional, emotional, sexual, cultural, family, and international domains, as well as, in political affairs. This evolution is of course ambiguous: the temporary contract is favoured by the system due to its greater flexibility, lower cost, and the creative turmoil of its accompanying motivations -all of these factors contribute to increased operativity" (Lyotard, 1984, p. 66).

The above quote makes clear that, funding of research would depend more and more on the research field's ability to demonstrate its capacity,to contribute to the system's performance by inventing new ideas, new games.

The role of IR in the context of Information Science as a research discipline should be considered from this perspective (see Wersig, 1993 for a similar view[48]). Viewed in the context of the current social and scientific setting, its legitimacy ultimately depends on the ability to transform and modulate existing paradigms to generate new ideas: "The function of the differential or imaginative or paralogical activity of the current pragmatics of science is to point out these metaprescriptives (science's "presuppositions") and to petition the players to accept different ones. The only legitimation that can make this kind of request admissible is that it will generate ideas, in other words, new statements" (Lyotard, p. 65).

5.7 IRS as encyclopedia

The semiotic analyses of chapters 4 and 5 treated the IRS as a code (see in particular sections 4.5.2 and 4.5.3) isomorphic in structure to the structural linguistics' model of langue (cf. 4.1.1, especially that of Hjelmslev's model described in 3.5.10. For the distinction between langue/parole, language-system/speech, see 3.5.5). This model is complemented with that of Peirce's model of signification or semiosis (cf. 3.5.2) to develop a synthetic model derived from the two major schools of thought in contemporary semiotics (cf. 4.1.2). It is however possible to expand on Peirce's ideas on semiosis and view IRS as a network of interconnected sememes (cf. 3.7.2), rather than a oppositional structure of signifier/signified, expression/content. If this structure (i.e. IR as a code) is a *dictionary* (see Warner, 90, p. 23), the network of sememes is an *encyclopedia*, speaking metaphorically.

5.7.1 Encyclopedia v. dictionary

When the componential approach to analysis of meaning was introduced in 3.7.2, it was noted that, this approach seeks to analyze meaning of linguistic units (lexemes) in terms of finite number of universal components (concepts, markers, properties, etc.). These are the so called semantic-primitives. A semantic-primitive is the most analytic (the simplest) concept that combines with other primitives to compose the meaning or sense of a given lexeme. To recite

[47] Lyotard use it in the text in the sense of *metaprescriptions*. Refer section 4.3.2.2 to see the context in which the term prescription is used by Lyotard.

[48] One should also note Froehlich, who seeks to establish a anti-positivist, anti-Cartesian, post-modernist foundation for Information Science (1989a), and information retrieval (the relevance criterion in particular) (1989b).

the example given in 3.7.2: lexeme /man/ is composed of semantic markers <<human + male + adult>> and/or <<rational mortal animal>>. This is the approach adopted in compiling dictionaries. Thus, my Oxford English Dictionary, defines /man/ as "an adult human male, esp. distinct from a woman or boy". A dictionary encapsulates the semantic competence of the speaker of a language by storing limited amount of information about a given lexical item (Eco, 1984, p. 49). The well known example of the componential analysis of the lexeme /bachelor/ by Katz and Fodor (figure 6.1) illustrates the format of an dictionary-like representation (in Eco, 1976, p. 97):

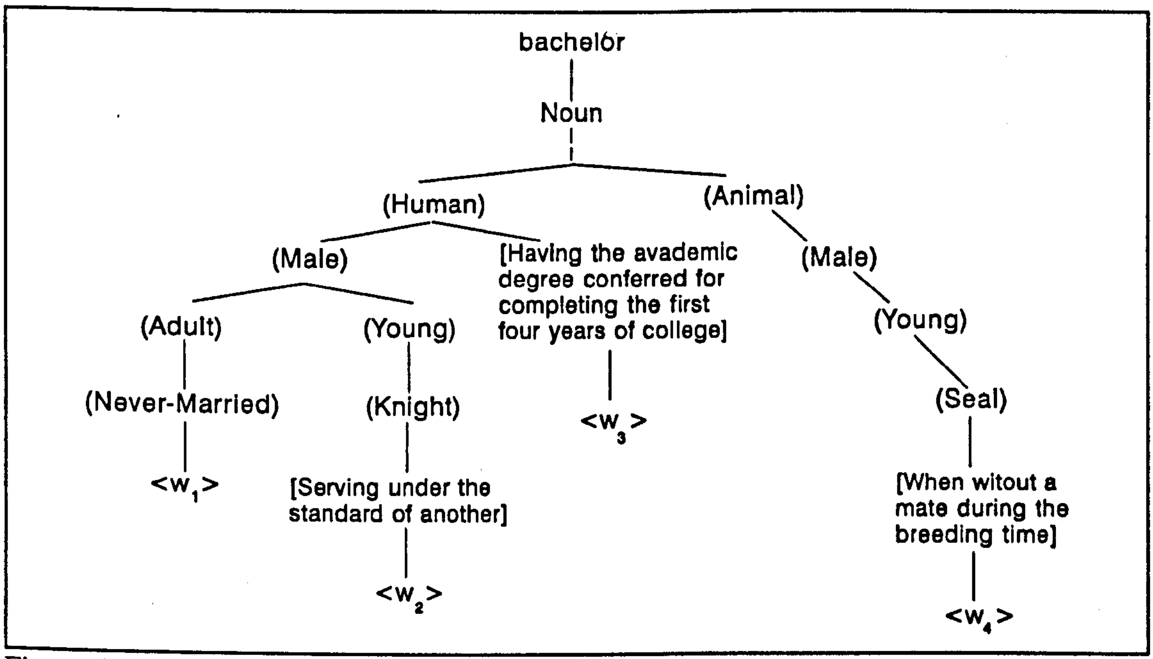

bachelбr
Noun
(Human)
(Animal)
(Male)
[Having the avademic degree conferred for completing the first four years of college]
(Male)
(Adult)
(Young)
(Young)
(Never-Married)
(Knight)
$<w_3>$
(Seal)
$<w_1>$
[Serving under the standard of another]
[When witout a mate during the breeding time]
$<w_2>$
$<w_4>$

Figure 5.7: The KF Model of Componential Analysis

In the above diagram sememe <<bachelor>>, which is the content of the lexeme /bachelor/, has the syntactic marker 'Noun'. It is also marked by semantic markers or semes shown in round brackets. The definitions in square brackets are called 'distinguishers'. Finally, each branch of the tree ends with a symbol in an angular bracket, called 'selection restrictions', prescribing necessary and sufficient conditions for the reading traced by a particular branch to combine with others (ibid, pp. 96-97).

It is sufficient for the purpose of this dissertation to observe the following two problems with the dictionary approach, as pointed out by Eco (1984, pp. 46-68): it is virtually impossible to know that: *a)* the primitives are indeed the simplest possible concepts that do not require any interpretation *b)* a restricted number of them is enough to compose unrestricted number of lexical items or words. In fact, the format of a dictionary assumes the equivalence of the defined lexical item to its definition. Thus the format of representation in a dictionary is $p = q$. However, even the simple case of the example of /man/ given above suggests that it is not straightforward to posit unequivocal equivalence between the definition and the defined. Each term in the dictionary definition of the lexeme /man/; <<rational>>, <<mortal>>, <<man>> require their definitions in turn. The markers in the above example are not absolute primitives, but are "... mere linguistic labels that cover more synthetic properties" (ibid, p. 68). Even the simplest dictionary entry requires some interpretation, thus, should be represented in the format: $p \supset q$ (cf. fig. 4.6 in 4.4.3).

Indeed in figure 5.7, the existence of definitions called distinguishers puts into question the purity (primitiveness) of the semantic markers (semes) employed for decomposing the sememe /bachelor/. The above model actually tries to explain the semes by more complex definitions than themselves. It is clear from this observation that, the above dictionary-like format demonstrates neither the existence of universal primitive concepts, nor the feasibility of production of compound concepts from restricted number of more analytic atomic concepts. It is worth to note here also that, the above model does not give the possible connotations that a term may have according to different contexts and circumstances in which it may appear. It may be sufficient for some purposes to contend with giving only the denotation of a term, but it is obvious that a full semantic analysis of meaning cannot do without the connotations associated with it[49].

The inferential model ($p \supset q$) on the other hand, in contrast to the dictionary-model, represents the content of a given lexical item by means of chains of interpretants in a process of unlimited semiosis (cf. 3.5.2). This is the format of the encyclopedia. In opposition to hierarchical, thus ordered and finite number of markers that makes up a dictionary, the encyclopedic model provides unordered, unrestricted set of markers[50] that codes the universe of shared (culturally known) knowledge. The format of the encyclopedia-like representation is discussed in more detail below.

5.7.2 The semantic space as encyclopedia

The following schema (fig. 5.8) from Eco (1976, p. 105) illustrates the format of the encyclopedic representation:

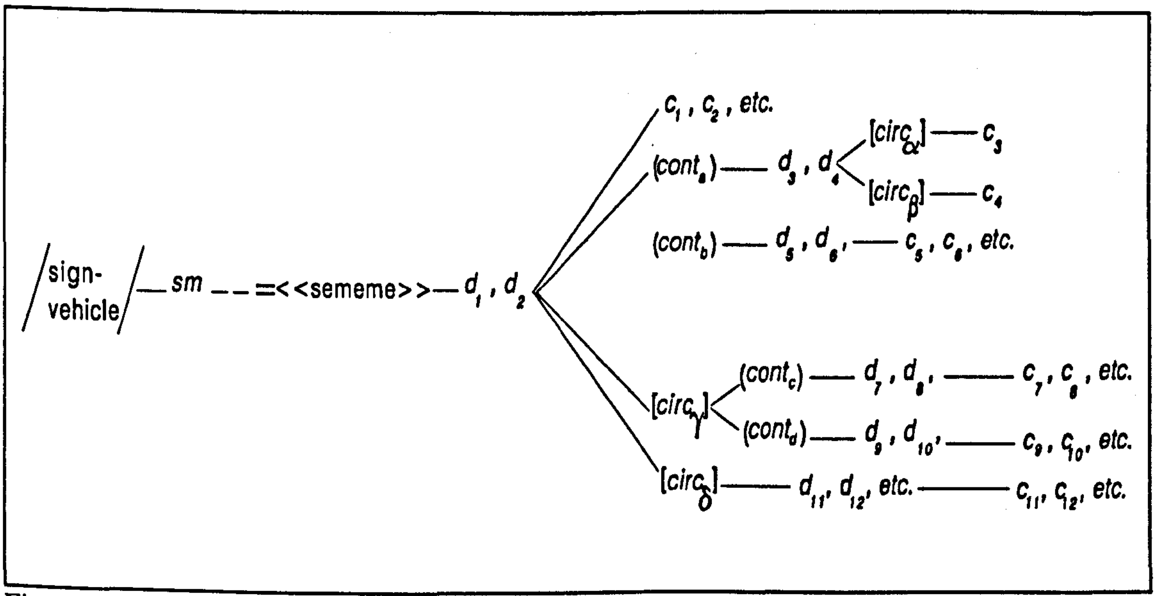


Figure 5.8: The Encyclopedic Model

[49] Thorough analysis/criticism of this model can be found in Eco (1976, pp. 96-105).

[50] These markers, in addition to lexemes or other linguistic units that act as interpretants (seme) (cf. 3.8) of other linguistic items (sememes) may include: frames, scripts and similar techniques that are used in AI research (see for a discussion of this: Eco, 1976, pp. 122-125; 1984, pp. 70-73).

In the above diagram, '*sms*' are the syntactic markers associated with the /sign-vehicle/ (i.e. with the expression). The denotations and connotations associated[51] with the <<sememe>> are shown with 'd_n' and 'c_n' respectively. The contextual selections, which are denoted by *(cont)* in the diagram, issue instructions of the type: "when *(cont$_a$)* is found, use the following *d*s, and *c*s, when the sememe in question is contextually associated with the sememe <<a>>" (ibid). The circumstantial selections, which are denoted by *[circ]* in the diagram issue instructions of the type: "when *[circ$_\alpha$]* use the following *d*s and *c*s, when the sign-vehicle corresponding to the sememe in question is circumstantially accompanied by the event or the object //α//[52], to be understood as the sign-vehicle belonging to another semiotic system or code" (ibid, pp. 105-106).

In this model, selection restrictions of figure 5.8 are eliminated by means of contextual and circumstantial selections, and distinguishers dissolve into a complex network of more analytic (simpler) semantic markers (semes). The most important feature of this representational model is that, each semantic unit used to analyze a sememe (i.e. acts as its interpretant or seme or semantic marker) is in turn becomes a sememe which is analyzed by other semantic units. This is obviously a model of infinite recursivity, referred to as unlimited semiosis elsewhere (cf. 3.5.2). Each of the *d*s and *c*s in figure 6.2 are therefore starting point of other trees, each of them constitute inside the tree a potential tree, a sort of embedded sememe generating its own tree, and so on *ad infinitum*. As Eco (ibid, pp. 122-125) remarks, this is a sort of representational model advocated in some AI work as well, for example by Quillian (1968).

The encyclopedic model of the semantic space constitutes an enormously complex ever changing (expanding, dying out, transforming, i.e. transitory/historical) network of interconnected sememes (interpretants), in which every sememe is ultimately connected with every other. This does not mean that, every connection is actually active in a given situation, nor known/established in a particular culture at a given point in time. It is rather the opposite. The contextual and circumstantial constraints make sure that only those which can be hypothesized (abduced, inferred, interpreted; cf. 5.6.1.2) successfully as the most plausible one in a given situation or setting, out of all possible connections are instantiated[53]. Furthermore, in a given culture at a given time, it is more likely that only some of the all possible links are known (collectively, i.e. culturally). However as mentioned in 3.8, the encyclopedic model conceives that new connections can be invented/discovered. This sort of newly proposed link, initially constituting a scientific or artistic invention, transforms into a metaphor or connotation once it is culturally accepted (conventionalized), and therefore known. This corresponds to linguistic creativity as it would be recalled from section 5.5.

Another important feature of the encyclopedic model is that, it only registers 'semiotic statements', that is, statements concerning cultural units. A cultural unit is a collectively shared knowledge. It is conventionally known within a particular culture. It is not to be confused with 'factual statements'. "Napoleon died on Saint Helena" was a factual statement on May 5 1821,

[51] Recall from 3.5.13 that, both denotations and connotations are cultural units, therefore should only be analyzed intensionally (i.e. without reference to objects or things out there in the external world) (cf. 3.5.14). The difference between the two is that of the primacy of coding, which is a cultural decision.

[52] The double slashes indicate the object used as a sign as it is the case in *ostensions* (cf. 5.3.2).

[53] This is the *pragmatics* of language use (cf. 3.5.4 and 3.9).

it is a semiotic statement today. An encyclopedia registers only semiotic statements, therefore, it is not concerned with truths, but what has been said about the truth (cf. 3.5.14)[54]. In this regard, the truth of the encyclopedic knowledge is questionable. An encyclopedia is also concerned with registering myths, legends, etc., in short, discourse about a subject matter, which can be (or believed to be) true, false, imaginary, and so on (cf. 3.7).

It may prove to be helpful to think such an encyclopedia in terms of the metaphor of the *rhizome* (Deleuze & Guattari, 1987). Rhizome, in contrast to root of a tree or to its branches which proceeds by multiplying itself by two (binary differentiation) constitutes a complex network of bulbs and tubers in which every point connects with every other. A rhizome can be broken off at any point and reconnected by following one of its lines. In this respect, it is impossible to produce global description of the whole rhizome. Only local descriptions which change in time could be provided. A rhizome is more of a map or a topographic chart than an image, such as a tree-structure, where the level of detail varies according to the point of view one adopts, and exact description is never possible. One can traverse two points on a map in a number of different ways (and sometimes indefinite ways), unlike a tree-structure which has always fixed connections all of which returning to the origin.

One should note that, the encyclopedic model of the semantic universe is only a regulative hypothesis, meaning that, an exact or complete description of the semantic universe cannot be possibly provided. It is a useful tool, in so far it isolates bits and pieces of the semantic space for practical purposes.

5.7.3 Document space as encyclopedia

It has been noted in 4.5.2 that, the meaning of the query statement (search statement) at denotation$_1$ is an abstract description of documents that match the form laid down by it. It is however relatively difficult to comprehend meaning in this sort of highly abstract terms. The metaphor of encyclopedia provides an alternative means of understanding the meaning of the query statement at this level of coding.

The relationship between the query statement and the documents in an IRS, according to the semiotic model developed in chapters 4 and 5, is that of between the signifier and the signified, or in Peirce's model of signification, between the sign (representamen) and its interpretant (cf. 3.5.1, 3.5.2).

If we loosely adopt the model of the encyclopedic representation to explain the relationship between the query terms and the documents, it becomes apparent that an index term can be conceived as a sememe and the documents indexed by it as its semes or interpretants (cf. 4.2.3). According to the encyclopedic model, it is of course possible to analyze semes in turn as sememes in which case the index term becomes one of the interpretants of the document in question.

This model helps us to conceptualize the meaning of the search statement at the first level of the model of fig. 5.5. The meaning at denotation$_1$ can be thought as the whole of the

[54] Whereas a factual statement such as a scientific statement regarding an object needs to be proven or tested against the facts in the world, a semiotic statement is accepted as true in a given culture (sometimes believed to be false, see below).

interpretants (sememes) associated with the expression, i.e. the query statement (cf. 4.5.2, 5.6.1.1 and 5.6.1.3).

To substantiate further the analogy between the encyclopedic model of representation of knowledge, and the representation of knowledge in an IRS, we can compare the instructions regarding contextual selections with that of indexing rules.

Indexing rules in an IRS (especially in a Boolean system) can be thought of as instructions of the following type: "when you find index-term$_a$ contextually associated (i.e., in the same query statement) with index-term$_b$, retrieve the following documents; d_1, d_2, ... , d_i" (cf. 5.6.1.3). The individual documents d_1 to d_i are the interpretants of the index terms (*index-term$_a$* and *index-term$_b$*) and vice versa.

However, if one wants to further pursue the analogy between the representation of sememes in the encyclopedic model, and the representation of documents in an IRS, it is necessary to look for equivalents of denotative and connotative markers (*d*s and *c*s in fig. 5.8), as well as circumstantial selections in an IRS.

One can hardly speak of circumstantial selections (cf. 5.7.2) in the context of documentary information retrieval situation for the obvious reasons. The sheer size of the document base in most operational systems precludes any possibility of incorporation of instructions of this sort in the IRS. However one can count efforts of the 'situational information retrieval' (Schamber, et al., 1990; Wilson, 1973) and 'sense-making theory' of Dervin (Dervin, 1977; Dervin & Nilan, 1986) as exactly this type of efforts . As it would be recalled from the last paragraph of section 5.7.2, the encyclopedic model is only a regulative hypothesis, and complete description of a semantic space as large as those represented in an average IRS cannot be possibly provided for the very simple reason that while one tries to arrest some portion of the semantic universe in some sort of formalism, it transforms itself already into a new structure with new connections between the semantic units. The new connections are the result of the interpretive processes taking place in a certain setting or situation. It is virtually impossible to abstract the meaning of things by extracting them from the stream of life that they are put in use in (Blair, 1992). There is this story by Borges which might be illuminating regarding to the above discussion, in which the emperor so obsessed with precision, partly because of his vanity of the extent of his empire, orders the entire population to produce the exact replica of the territory ruled by him, which eventually drives the empire into bankruptcy.

The model of encyclopedia discussed in 5.7.2 has also denotative and connotative markers (fig. 5.8), of which, the connotative markers depends on the denotative ones in terms of the order of signification as a result of the cultural conventions of coding (cf. 3.5.13). To fully explore the isomorphism of both models, it is necessary to discuss whether or not a similar order exists in an IRS. It is difficult to envisage to impose an order of retrieval in the document retrieval situation in which the importance or meaning or use of a document depends on the prior retrieval (or knowledge of it by the user) of another document. There exists no readily available coding rule(s) which tells us the order of retrieval in terms of dependence of signification on the inter-document relations. However the 'cluster hypothesis' (van Rijsbergen & Sparck Jones, 1973) and more particularly Goffman's (1969) proposals, as pointed out by Robertson (1977a), are methods based on dependence between documents and therefore could be counted as efforts of this sort. One can conceive the method of retrieving documents in pairs as proposed by Wiesner (1988) in terms of the denotative/connotative markers as well.

It can therefore be legitimately expected from this exploration of the encyclopedic model within the structure of IRS that, both the order of retrieval and the situation in which the interpretation of the texts is taking place have an impact on the significance or use of the document to the user of the IRS. These conclusions have already been drawn by many others (e.g. Dervin & Nilan, 1986; Saracevic, 1975; Schamber et al., 1990). To admit that both of these relations are difficult, even impossible to implement in a structure like that of an IRS, does not preclude the possibility of drawing an analogy between an encyclopedia and a document retrieval system.

Regardless of these difficulties, the encyclopedia metaphor gives useful insights to the retrieval situation and deserves more attention. It is for example quite a good exercise to think the relation between the query and the documents in terms of the infinite recursivity of the encyclopedic model (5.7.2). Documentary information retrieval situation has already been discussed in chapters 4 and 5 in terms of unlimited semiosis (see especially 4.2.3, 4.3.2.1 and 5.4.2). The encyclopedic model makes possible to think about index terms and documents as semantic units interpreting and explaining each other by perpetually pointing (signalling) one another, and continuously referring the user to another document or an index term (cf. 6.3.2) without ever closing on itself as a stable system.

Finally, it might be interesting to compare encyclopedia as an artefact to information retrieval mechanisms. It has already been noted that, meaning at the denotative level (denotation$_1$; cf. 5.6.1.1) can be thought in terms of the whole (aggregation) of the semes that interprets a given sememe in a universe represented by a knowledge model such as an encyclopedia. The same sort of coding (i.e. one that is called denotation$_1$ in 5.6.1.1) can be found in an ordinary encyclopedia (artefact) when one thinks an encyclopedia as a search device, that is an index. The encyclopedic organization of entries in alphabetical order certainly constitutes (practically) an index. Similar to an information retrieval mechanism, index of an encyclopedia correlates an index entry with a text. Some of these index entries may also be homographs in which case one can say that (in principle at least) an index entry is correlated with a number of texts. This makes the analogy between an IRS and an encyclopedia easier to conceive. Texts (semes) that are directly correlated with an index entry (sememe) in an encyclopedia can, therefore, be conceived as the meaning(s) of the index term at denotation$_1$ (cf. 5.6.1).

The connotative level (cf. 5.6.2) in an encyclopedia can then be interpreted as various ways of readings of the individual semes connected with the sememe in question. The two sub-levels (denotation$_2$ and connotation$_2$) of the connotative level of section 5.6.2 can be defined in terms of the semes (interpretants) associated by the user with the text of the encyclopedic entry that s/he is reading. To restrict oneself in associating a given text with a small number of known semes can be thought of similar to the labour that involves recognition of meaning(s) of signs (denotation$_2$, cf. 5.6.2.1), on the other hand, a free associative reading which is not restricted to an *a priori* set of known semes can be thought as an inventive or interpretative reading (connotation$_2$, cf. 5.6.2.3).

Chapter 6
Application of the Semiotic Theory to Systems Design Problem

The objective of this chapter is to apply the semiotic theory of information retrieval developed in chapters 4 and 5 to Information Retrieval Systems design practice. In doing so, the semiotic model developed so far is reviewed and complemented as necessary with further analyses of the retrieval situation. Section 6.1 presents an analysis of the design philosophy and underlying retrieval model of the Okapi system which is used as a platform to apply and test some of the ideas developed in this project. In section 6.2, interaction in Okapi information retrieval system is analyzed in terms of the semiotic concepts and tools of chapters 4, 5 and 6. Section 6.3 leads us to the discussion of the IRS design objectives and criteria within the constraints of this project. The final section of the chapter (6.4) summarizes the discussion and gives an overview of the systems design and evaluation objectives of the project.

6.1 Okapi experimental information retrieval system

The objective of this section is to discuss the underlying philosophy of design of the Okapi retrieval system from the perspective of the semiotic theory of information retrieval developed in the preceding chapters. Okapi retrieval system provided the platform for applying some of the ideas discussed in this dissertation (see chapter 7). The criticism of the current operational system from the perspective adopted in this dissertation helped to clarify the design objectives (see 6.3) of the knowledge-based systems which are described in the next chapter.

The next section (6.1.1) discusses the design philosophy of the Okapi system. In the subsequent two sections the underlying retrieval model are examined (6.1.2) and criticised (6.1.3).

6.1.1 The retrieval model

Okapi is an operational interactive information retrieval system based on a probabilistic retrieval model which makes use of relevance feedback information. A detailed description of the system is given in 7.2. In this section the principle upon which the retrieval model in Okapi is based, discussed.

The retrieval model employed in Okapi is based on the 'relevance weighting theory' of Robertson and Sparck Jones (1976). It has been further developed by Robertson (1990). A detailed discussion of this model is presented in 7.2.2. In the following paragraphs more general features of the model is examined.

There may be different reasons for weighting the terms in a search statement (Robertson, 1990; Robertson & Sparck Jones, 1976). The basic rationale behind the 'weighting model' of Okapi as developed by Robertson and Sparck Jones is that, a document retrieval system should rank the documents presented to the user in the order of their probability of relevance to the user's need (Robertson, 1990; 1977a). This principle is formulated by Cooper as the 'Probability Ranking Principle' (PRP): "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data" (in Robertson, 1977a, p. 295).

The objective of the PRP is to rank documents according to their probability of relevance to the user. As stated by Robertson & Belkin (1978a, p. 96) 'probability' enters the IR picture because of the discrepancy between stated request for information (query statement) and the latent need for information underlying the stated request, as well as, the system's ability to make inferences (about the need on the basis of supplied data, i.e. the stated request). If the needs could be stated exactly and completely, and the indexing of texts were complete and exact, the probability concept would not arise. It is the axiomatic of PRP that the discrepancy between needs and requests cannot be eliminated, therefore, there is always some probabilistic element in IR (ibid).

However to arrive a simple ranking rule several other simplifying assumptions are needed. The most important of them are: *i)* the relevance judgement is dichotomous, i.e. a document is either relevant to the need or not (ibid, p. 94), *ii)* the relevance of a document to a request is independent of the other documents in the database (Robertson, 1977a). Even then one does not necessarily arrive a simple ranking rule. There are cases where PRP can lead to non-optimum results (Robertson & Belkin, 1978a, p. 96).

One needs also to distinguish ranking according to probability of relevance from ranking according to 'degree of relevance' (Robertson & Belkin, 1978a). The former assumes a dichotomous relevance variable. The latter hypothesises that different documents satisfy the user's information need to different degrees. It assumes that the relevance of a document to a need is measured along a continuous scale varying from a non-relevance to high-relevance (Robertson & Belkin, 1978a, p. 94). A formal model of degree of relevance is given in (Robertson, 1977b). Another reason for wanting to rank is therefore, to present the user the documents with higher degree of relevance first (Robertson & Belkin, 1978a, p. 94). There exists however, no formal method of ranking according to degree of relevance[1] and in practice the two principles are often confused. In evaluation experiments for instance, when scale of relevance is introduced to the measurement of degree of relevance, it is most often transformed into a dichotomous variable in the later stages of analysis (ibid, p. 95).

It has been hypothesized by Robertson & Belkin (1978a) that while the probabilistic concerns enter into IR mainly because the discrepancy between stated needs and latent need or ASK, the degree of relevance enter into play because of the relationship between text and need, and nothing to do directly with the request nor with the IR mechanism (ibid, p. 96). The relationship between the two principles is a complicated one and there exists no model that combines the both views (ibid, p. 98).

6.1.2 The PRP as transmission of information

In the preceding section the principle behind the retrieval model of Okapi is described. In this section some of the important assumptions underlying this principle are discussed with a view

[1] Goffman's method (1969) being the exception. The 'cluster hypothesis' has some potential in this regard (Robertson, 1977a).

to relate them to some of the themes developed in chapters 4 and 5.

The following is a summary (though not a complete list) of the explicit assumptions used in the formulation of the PRP:

a) the relevance of a document to the user is independent of the other documents in the database (Robertson, 1977a; Robertson & Sparck Jones, 1976)

b) the document is either relevant to the user or not (Robertson, 1977a; Robertson & Belkin, 1978a; Robertson & Sparck Jones, 1976)

c) the ideal response of a IRS is to retrieve all the relevant documents and none of the non-relevant ones. Therefore, there exists an exact set of documents that will be judged relevant by the enquirer, i.e. that perfectly satisfies the user (Robertson & Belkin, 1978a, p. 94)

d) there is always some inevitable discrepancy between the stated request and the actual information need of the enquirer (Robertson & Belkin, 1978a, p. 96)

From the above four assumptions follow a number of mostly implicit assumptions regarding the information retrieval situation which are related closely to the above given list. The important ones from the point of view of the discussion of this section are given below:

i) the user interact with the IRS, because of some lack

ii) this lack is manifestation of a problem or a task that can be described adequately in terms of definite boundaries, therefore can be solved optimally

iii) the information needs of the inquirer and the texts in the database exists as separate (independent) entities in the world

iv) the inquirer basically has one query, or if there are more than one, they can be safely treated as independent queries

Starting with the first point above, it is explicitly stated by several authors that users consult IRS because of some underlying need to resolve a problem situation (see e.g. Belkin, 1980; Robertson & Belkin, 1978a). What is not explicitly stated though is that, it is assumed that this is a pre-defined problem that can be solved here and now (cf. 5.5.5; 5.6.2.3). In other words, it is assumed that there is some information lacking from the user's knowledge of the problem and the purpose of interaction is to fill in this lack or gap in order to solve the problem[2]. To make clear what it implicitly suggests, what it does not say should be explicated: It does not say for instant, that user might have something in excess (a surplus value)[3], that s/he is trying to produce new information, new knowledge, a new problem, in short a new move (cf. 5.5 and

[2] Therefore, *implicitly* assuming that the problem can be solved if there is enough information regarding the problem.

[3] From the analysis of the rule-changing activities in IR (cf. 5.5) it can be argued that in the cases of interaction with an IRS where the priority is the production of new statements and moves (cf. 5.5. and 5.6.2.3) rather than consumption of knowledge (cf. 5.6.2.1 and 5.6.2.2) the user has more knowledge than say the subject of didactics (cf. 5.6.2.2.) that *overflows* the schema of relevant/not-relevant (cf. 6.4.2), i.e. recognition of meaning.

5.6.2.3; 5.6.2.4).

This discussion should become clearer if it is taken in connection with the point *'ii'* above. The implicit assumption of the PRP (along with the others, e.g. the ASK hypothesis) is that, the user is consulting the IRS for one reason only; s/he has a problem to solve which can be formulated as choosing among the alternatives, the best, i.e. the optimum solution. This in its turn implicitly assumes that, a problem space has already been defined, and what is needed is to search the space for the optimum answer (see Winograd & Flores, 1986).

The PRP therefore implicitly assumes consumption of information in contrast to its production (cf. 5.5.5 and 5.6.2.3). It implicitly encourages consumption and transmission of information at the expense of innovative moves that creates new connections and new ideas. It encourages this because it assumes that, the user has a pre-defined problem and the documents are judged according to their pertinence to the already formulated problem. This is as would be recalled from 5.6.2.1 is akin to recognition of ready-made expressions.

Furthermore, the documents are retrieved and judged one-by-one (see point '*a)*' above) as isolated discrete units (cf. intertextuality in 4.2.3). This further amplify the tendency in the PRP for transmission/consumption of information. A set oriented strategy such as Boolean retrieval for example, does not suffer from this tendency of closing on itself (an illuminating comparison of set-oriented retrieval versus document-by-document view can be found in Bookstein, 1989).

An important feature of the PRP, as discussed above, is the implicit suggestion that retrieval behaviour is exclusively motivated by some need, i.e. lack. This implicit suggestion is made by discourse such as that can be found in Belkin, 1980 and Robertson & Belkin, 1978a. This discourse on information retrieval in its turn suggests that this lack should be filled in by the transmission of information, relying on the metaphor of some kind of flowing fluid that fills in the vacuum (the 'mystical fluid' metaphor as remarked in Liebenau & Backhouse, 1990), amplifying the already strong tendency of consumption inherent in the model.

The idea of lack dominates the Western philosophical tradition for a long time and one should be aware of its short comings and the way in which it is used for various ideological purposes (Deleuze & Guattari, 1977). As it has been discussed briefly in 5.5.3 and 5.6.2.3, the lack and the related idea of transmission of information has ideological underpinnings, which needs to be acknowledged[4] and reflected upon.

The next section presents a critical view of the 'transmissional' models in IR, i.e. models that advocate albeit implicitly reproduction/transmission of information with a particular strong focus on the PRP, which will prepare the way for the discussion of the design objectives of the knowledge-based systems in 6.3. In doing so, the points '*iii)*' and '*iv)*' posed at the beginning of the section will be taken up and examined.

[4] Any signifying practice is performed against an ideological background (Barthes, 1972). It has not been suggested here that they should be avoided (as a matter of fact ideology cannot be avoided). However, it needs to be explicitly acknowledged wherever possible. It is a political (and sometimes practical) matter whether or not one then decides to address it.

6.1.3 Criticism of the transmissional models in IR

The PRP is an important statement, opening the way for the relevance revolution (Robertson & Hancock-Beaulieu, M.M., 1992). In fact the relevance revolution had prepared the way for what is usually called the 'cognitive paradigm'. The crucial turn taken by the relevance revolution is to locate the relevance criterion outside the retrieval system: "In this paper, I will take the relevance (or usefulness, or user satisfaction) to be a basic, dichotomous criterion variable, defined outside the system itself" (Robertson, 1977a). This move made possible what is generally known as the relevance feedback mechanism in IR. From there on, it was not a big leap to take the user as a part of the IRS.

One can therefore legitimately expect much common ground between the two important discursive turns of the last two decades in the IR research. These two turns, i.e. the relevance feedback mechanism and the user-oriented studies or information-seeking behaviour studies however carried with themselves, as discussed in the preceding section, the implicit suggestion that IR is exclusively about transmission of information to solve pre-defined (i.e. known) problems.

This implicit assumption presupposes another assumption: documents and information needs are two qualitatively distinct (heterogeneous) entities, the former a linguistic system, the latter a cognitive or mental phenomenon.

The analysis of chapters 4 and 5 showed however that, the separation of documents from other signifying systems is unfounded. This becomes apparent when the problem is analyzed by the appropriate tools developed by the disciplines that study signifying practices such as, semiotics, philosophy of language, literary criticism, cultural studies, anthropology, and so on (see, in particular sections 4.2.3 and 4.3.2.1).

This discursive turn has accompanied with another closely related move which takes the relevance criterion as a uni-dimensional variable, measured either on a dichotomous scale (as in the case with the PRP, cf. 6.2.2) or on a continuous scale.

This view is closely echoed by the move towards taking granted that the users search for topics with clearly demarcated boundaries, so that, although an inquirer might have more than one topic in mind, the spill-over between topics can be neglected, and each query can be treated as an individual distinct need. This implicitly makes the assumption that what is referred as intertextuality in the previous pages is either negligible or non-existent.

Thus, the cognitive paradigm and PRP/relevance feedback oriented approaches feed on each other's implicit assumptions symbiotically, and provide support for both paradigm's common claim for the nature of the interaction between knowledge, knowing, knower, meaning and information.

Both paradigms, however, exclude the cases where interaction with the IRS cannot be reduced to (rational/logical) search for answers to known questions, thus, to transmission of information (cf. 5.6.2.3). Furthermore, one should not loose the sight the conditions under which the above approaches to information (i.e. information transfer oriented approaches) are justified (legitimized) (cf. 5.5.3).

To conclude the criticism of the transmissional models of IR, the following common characteristics of the various different approaches can be noted:

i) The objective of the retrieval activity is to find the missing information necessary to perform a certain task (to solve a problem).

ii) The task engaged in, that is the problem to be solved, is known *a priori*, if not by the inquirer herself, by the experts in the field.

iii) As the problem is *a priori* known, the information needed to solve it known *a priori* as well.

The above points amount to say that, the purpose qf IR is to match the *a priori* known document set with the *a priori* defined problem set.

The criticism of this view has been presented in detail in chapters 4 and 5. It is suffice to note here that, there are cases in which this does not hold true. In fact, this is an exceptional case, where the user (the agent) is (or assumes the role of) the student, i.e. the subject of the didactics situation. It has noted by various authors that the model of rational search for an answer of the decision theory cannot be taken for granted in whole range of real life problems (see e.g. O'Connor, 1993; Winograd & Flores, 1986; Rapoport, 1966).

The next section (6.2) articulates the above criticism of transmissional models of IR in terms of the systems design practice.

6.2 System Design from a Semiotic Perspective

The objective of this section is to articulate the main features of the semiotic view of the IR process developed in the preceding chapters (6.2.1) and analyze the interaction in Okapi from the perspective of the semiotic view developed so far (6.2.2) with the objective of deriving criteria for IRS design which is done in the subsequent sections 6.3 and 6.4.

6.2.1 Main features of the semiotic model of IR

The objective of this section is to summarize the main characteristics of the documentary information retrieval process articulated by the semiotic analysis of the preceding pages. The various sections below discusses the related features of the semiotic model of the IR situation.

6.2.1.1 Coding in IRS

Semiotic analysis of the IR process reveals that there are two main levels of coding in IRS (cf. 4.5.2, 5.4). The relationship between these two levels sheds light into the relation between the user's query and the documents in the database which is the main problematic that concern many IR research work.

The fundamental question that any IR research must face is the nature of the human querying of the retrieval system. More precisely the question is: "*what makes the user to come to consult an IRS in the first place?*". The answer to this from a semiotic perspective is given in 4.3.2 in terms of the tripartite model of 'social text - science paradigms - scientific communication channels'. This model is further fleshed out in 5.5 with the introduction of the concepts of language games and in particular in terms of the pragmatics of the scientific knowledge, i.e. the 'research game' and the 'teaching game'.

The most important feature of this model is that it replaces the dichotomy of mental models/linguistic structures which is abundant in IR theory making by inserting the social subject (that is the user of IRS) into the social systems that shape her informational or otherwise behaviour.

In the bi-layered IRS model of chapters 4 and 5, the first layer called the denotative layer (denotation$_1$) has a particularly important role to play in the functioning of IRS: the complex process of retrieval of documents relies on an initial mechanism of describing the topological characteristics of the documents to be retrieved. To put this in another way, the expression tokens in IR is produced according to content-models or types (cf. 4.2.3, 4.3.2.3, 4.5.2).

One can in this regard note the similarity between the production at this stage of IR (i.e. denotation$_1$) and other sort of sign-vehicles governed in their production by ratio difficilis (cf. 4.2.3), such as, photographic and televised images.

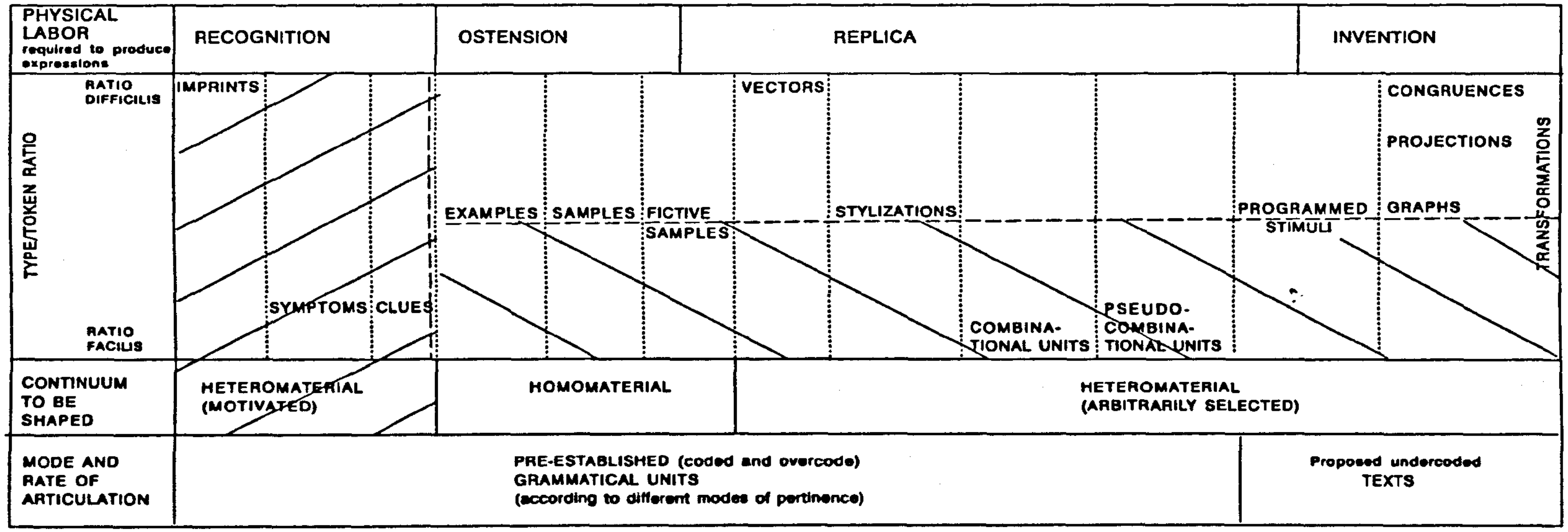
It is important to note here therefore the radical contrast between the search statement (expression) in IR and natural language discourse of the documents in terms of the structures that govern their generation. The fundamental differences between signs and symbols (images) has been discussed in 4.5.3. It is revealing in this respect to re-examine the chart given in figure 5.1 in terms of sign production capabilities allowed by IR mechanisms.

There are two main structural limitations that an IRS imposes on the selection of the 'mode of sign production' at denotation$_1$. The first and far more important is the type/token ratio. As, this is ratio difficilis in IRS (cf. 4.2.3), the mode of sign productions left below the line dividing the diagram into two halves in the middle are structurally impossible to utilize in IR. The modes of production lying in the middle of the diagram are border cases that share properties of both varieties of type/token ratios.

The other constraint imposed by the structure of the IRS is to do with the relation between the physical continuum shaped by the expression and the physical material that its possible referent might have been made of (the 'continuum to be shaped' parameter in fig. 5.1 in section 5.2). It is clear that in IR context only the category of 'motivated-heteromaterial' can be crossed-out outright. The resulting structural possibilities in terms of expressiveness of an IRS is shown in fig. 6.1 below.

It is apparent from fig. 6.1 that one cannot expect the sort of expressive possibilities that is found in natural language discourse ('combinational units' in *fig. 6.1*; cf. 5.3.3) from an IR mechanism (see also the discussion in 4.5.3). This analysis reveals that if a natural language discourse mechanism such as a question answering system is to be designed, a totally different structural organization of expression and content planes is needed. In other words, one cannot possibly expect natural language discourse type of interaction from a system designed for document retrieval[5].

[5] One can of course possibly design an user-system dialogue sub-system in an IRS in which the query statement finally submitted to the system is generated as a result of a dialogue between the user and the system, which may resemble or have the same properties of a natural language discourse. However, beyond this sub-system, the relation between the actual search statement submitted to the system and its contents would still be governed by ratio difficilis.



PHYSICAL LABOR required to produce expressions

RECOGNITION

OSTENSION

REPLICA

INVENTION

TYPE/TOKEN RATIO

RATIO DIFFICILIS

RATIO FACILIS

IMPRINTS

SYMPTOMS

CLUES

EXAMPLES

SAMPLES

FICTIVE SAMPLES

VECTORS

STYLIZATIONS

COMBINATIONAL UNITS

PSEUDO-COMBINATIONAL UNITS

PROGRAMMED STIMULI

CONGRUENCES

PROJECTIONS

GRAPHS

TRANSFORMATIONS

CONTINUUM TO BE SHAPED

HETEROMATERIAL (MOTIVATED)

HOMOMATERIAL

HETEROMATERIAL (ARBITRARILY SELECTED)

MODE AND RATE OF ARTICULATION

PRE-ESTABLISHED (coded and overcode) GRAMMATICAL UNITS (according to different modes of pertinence)

Proposed undercoded TEXTS

FIGURE 6.1: Expressive possibilities of IRS (after Eco, 1976, p. 218)

However, when the query statement is replaced by the system with some retrieved documents, the user enters a different mode of interaction, that of natural language discourse (cf. 4.5.2). At this level one can speak of two radically different types of language games in play (cf. 5.6): denotation (denotation$_2$) and prescription (connotation$_2$). The relationship between the two levels, i.e. denotation$_1$ and denotation$_2$ and/or connotation$_2$, is vital in understanding the interaction of the user with the system in IR. The user at denotation$_1$ rarely produces his query separately from his expectations in terms of either denotation$_2$ or connotation$_2$ or both (cf. 5.6.1.3, especially footnote 13).

In producing the query statement, the user must take into account the inferential instructions of the system. The inference in IR is located on the user side. If the inferential instructions of the system are complex (cf. 5.6.1.3), it is harder for the user to make sense of the system. However, the important point here is that, this hardship is neither psychological nor subjective. This is a very important point to note. The assumption that this hardship is due to human incapability to process complex information can only be justified if one assumes that the purpose of the IR is to find the best or right answer (i.e. already formulated or known answer) to an already formulated or known question (cf. 5.6.2.3, 6.2.2 and 6.2.3). This does not take into account the research pragmatics (cf. 5.5.4, 5.6.2.4). The ability to control both the course of the search and its outcome (prediction of the result) are part of the user's competence of system use (i.e. the user's competence of performance, which is similar to competence of performing natural language acts, cf. 3.9, 5.5.5). In this regard, decipherability of inferential instructions are related to the user's cognitive abilities. As when one is not competent in a natural language, one cannot think in that language, similarly when the rules of the system are incomprehensible to the user, the user cannot think, research, learn, in short work with that system. Comprehension of the system's inferential instructions and formulation of the search statement are all part of user's competence of the retrieval situation.

Regardless of whether or not the system provides the best answer, the act of retrieval cannot be considered fulfilled without the user's responsibility to initiate, control and evaluate the process. These responsibilities and the associated responsibility (and power) to think cannot be delegated to a system/machine without a fundamental decision concerning the social consequences of the whole process which is itself a political matter.

6.2.1.2 Languages games in IR

The process of document retrieval cannot be considered as a homogeneous linear flow of information from the addresser to the addressee and back. It is rather a heterogeneous multi-levelled, multi-code discursive interaction (cf. 5.6). Two distinct informational games are identified in the context of IRS: denotation and prescription (cf. 5.5.5). The distinction between the two has not been made generally in the IR literature (with some exception, e.g. O'Connor, 1993).

The denotative game in the context of IR involves recognition of an information need and the corresponding answer or documents (cf. 5.6.2.1) which are usually linked together by means of the concept of relevance (cf. 5.6.2.2). The denotative game is about transmission of known facts (i.e. information) from an authority of some sort (the expert, the database, etc.) to the subject of didactics, i.e. the student. This is a case of rule-following (rule-governed production; cf. 5.5.1). The student follows the rules of the code[6] at both levels (5.5.1): denotation$_1$ and

[6] That is the IRS. See 4.3.2 for the functioning of IRS as a code.

denotation$_2$ (see fig. 6.2 below).

The prescriptive game includes production of new statements by establishing novel linkages between documents (cf. 5.5.5 and 5.6.2.3). This is about production of new knowledge. These inventive (prescriptive) productions in IR can be considered as rule-changing or rule-making activities (5.5.2 and 5.5.5). The rule-changing activities in IR can be subsumed under two main headings (fig. 6.2): moderate and radical inventions.

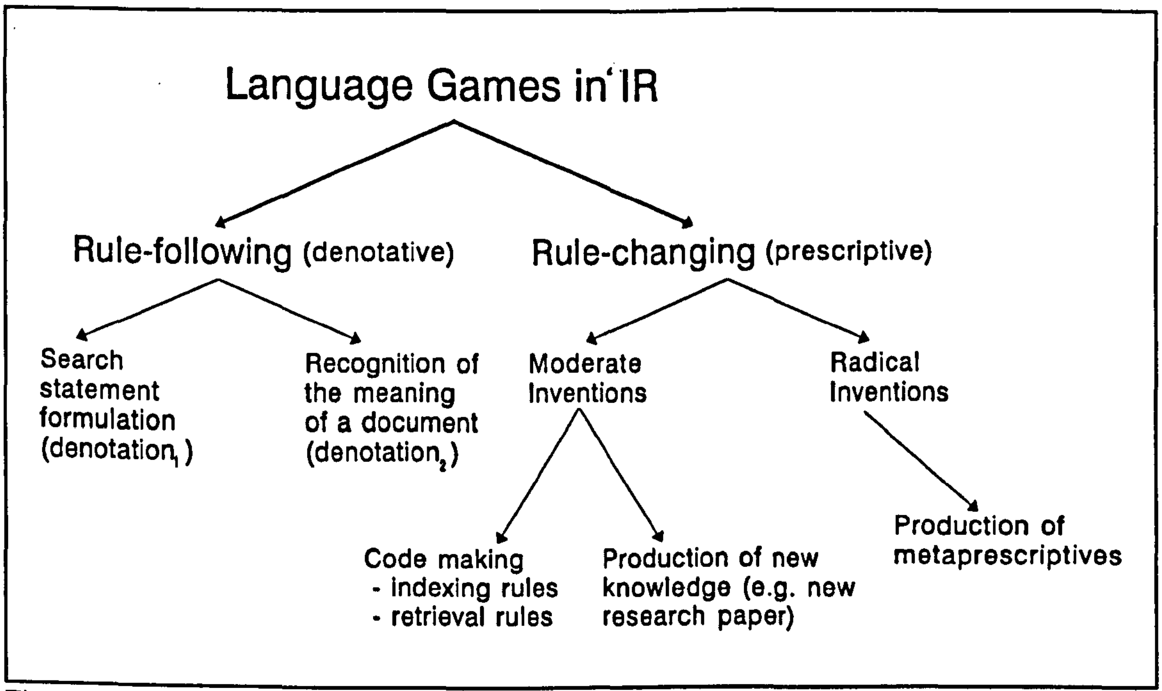

Language Games in IR
Rule-following (denotative)
Rule-changing (prescriptive)
Search statement formulation (denotation$_1$)
Recognition of the meaning of a document (denotation$_2$)
Moderate Inventions
Radical Inventions
Code making
- indexing rules
- retrieval rules
Production of new knowledge (e.g. new research paper)
Production of metaprescriptives

Figure 6.2: Language Games in IR

The decision on the rules of indexing and retrieval constitutes system design activity by the experts of domain and/or IR, and referred to as code-making (cf. 5.5.1 and 5.5.2). System effectiveness evaluation tests under laboratory conditions, often involving parameters related to rules of indexing and retrieval, are this sort of activities (cf. 5.5.1 and 5.6.2.3). Code-making activity in its capacity of affecting the content representations of the retrieval systems can considered as an example of moderate inventions in IR. Moderate inventions involves directly the content model of the system (cf. 4.3.2.3 and 5.5.5). Research publication (i.e. production of new knowledge) which eventually ends up in some sort of indexing/abstracting service for retrieval can be considered as another type of moderate invention which changes the content model of the code.

Radical inventions in IR context deal with the rules (metaprescriptives) governing the admissibility of moves (statements) in a given paradigm, therefore, involve the experts in the field (cf. 4.3.2.2 and 5.5.5).

Certain types of rule-changing activities, such as those involve production of new knowledge (5.6.2.3) can be thought to be akin to what O'Connor (1993) calls 'searching without a topic' or 'browsing'. Browsing in contrast to 'grazing', which is similar to what is called as 'transmission or reproduction of information' in this dissertation, consists "... of a wide spectrum of idiosyncratic processes for searching, sampling, and evaluating of documents ..." (ibid). It is described by O'Connor quoting Overage and Harman as "willingness of the scholar to search in a literature not obviously relevant" (ibid). This type of search resembles to dialogical interaction between two people or to everyday conversation without a well defined topic as the object of the exchange, while the pragmatics of information transfer is a relatively straightforward cooperative transaction between the sender and the addressee. Evaluation of the interaction is rather complicated in the case of rule-changing activities in IR, and examined more thoroughly in the next section.

6.2.1.3 Relevance

The idea of relevance is one of the most debated topics in IR that has yielded little results so far (Robertson & Hancock-Beaulieu, M.M., 1992). The failure to explicate the concept clearly stems from the failure to distinguish between the two types of IR pragmatics named above as transmission/reproduction of information and production of new knowledge. As remarked in 5.5 and 5.6, relevance judgement as an *instantaneous* decision about the relation between an information need and its referent (the documents) is simply not pertinent in IR pragmatics concerned with production of new knowledge.

Evaluation is an integral part of any type of searching. The important distinction between evaluation in the case of search for known facts (information transfer, the didactics situation) and in the case of inventive labour should however be clearly noted.

Evaluation in the latter case is a much more elaborate and complex process compared to former's relatively simple match between a pre-determined problem and its solution, a ready-made expression and its coded content[7] (cf. 5.6.2.1). As O'Connor (1993) remarks in the pragmatics of creation of new knowledge, creation of new linkages between documents, evaluation takes into account the *possible* value of an *unlikely item*. This sort of evaluation requires considerable studiousness, as identification of a bad item may not be self-evident: "... simply throwing out the bad are not entirely self-evident and they may well bear little resemblance to the formal methodologies of any relevant discipline ... A linkage may not itself prove valid yet show the way to another which might" (ibid p. 226).

The main thrust of this dissertation can be paraphrased as: **creating a new theoretical object of research in IR by separating what is *usually* in *practice* intertwined processes of denotation (description) and prescription (action).**

It has already noted in 5.5.5 that, the two labour are much closely related in actual IR pragmatics to be treated separately, although, there may be cases in which the distinction becomes more clear cut. However, the real importance of this somehow artificial separation rests in the theoretical and methodological framework it provides for extending the boundary of the system in the pragmatics of IRS design to the retrieval activity prior to approaching the IRS with

[7] Even in the case of simple information transfer (the case of didactics) user's relevance judgements are normally affected by factors other than the documents' subject matters, such as the style of the author, accessibility of the material, readability of the material and many others. However the point here (at least from the point of view of subject access) is that, the correspondence (or relevance) between a query and the documents can be determined objectively by the experts in the domain. This is not the case at all when the labour involved is the production of new knowledge (cf. Karamuftuoglu, in press).

a set of prescribed relevance criteria corresponding a vague or well defined problem state. This entails in its turn, extending the research problematic in IR to the institutional (5.6.2.4, 5.5.3) and societal (5.5.4) levels.

It has been articulated in 5.5.5 and other relevant sections of chapter 5 that, there are cases of retrieval activity, such as the research pragmatics, in which the relevant features of documents are not a priori given, but prescribed in the course of the research activity. The prescribed relevance criteria may not be known at the beginning of the research or even until to the very end of the whole process. The relevance criteria may be completely new as in the case of radical inventions (5.5.2, 5.5.5) and paralogical practices (5.5.4), or it may be partly or wholly recognizable as one of those already known problem classes (5.5.5).

Whatever the nature of the relevance criterion is, it should be defined by the user, either creatively, as in the case of research or, passively as in the case of didactics. It was hypothesized in the preceding chapters that it was the implicit assumption of prevailing IRS paradigms that the relevance concept involves recognition of intrinsic properties of documents, therefore, they are not constructed or invented by the user, but are out there to be discovered by the inquirer.

The challenge of the semiotic view of IR is to re-conceptualize the relevance criterion by first arguing that, it is a construction rather than an intrinsic property, and then bringing to the fore the underlying institutional and societal processes (4.3) in the construction process. The IRS design pragmatics therefore becomes a political process where a decision has to be made between transmission or invention oriented views of the process (5.6.2.3). The relevance concept then takes the form of either *recognition of relevant properties of documents* or *invention of new linkages of relevance*, accordingly.

6.2.1.4 Text

IR ultimately is about texts. However, text should be understood in its broadest possible sense, i.e. all signifying practices -- verbal or not (4.2.3). Furthermore, each text is a meeting place of several others which refer to each other inextricably (4.3.2.1, 6.1.2).

Retrieval of texts is not only the main objective of the IR process, it is at its inception as well: as it would be recalled from 5.6.1.1 it was remarked that in IR, "one starts from the text", rather than the other way around, from some mental or cognitive state. The importance of this statement should become clear when retrieval activities usually subsumed under the umbrella term of browsing is examined.

O'Connor (1993) makes a distinction between the two senses of the term 'searching' by differentiating browsing from grazing (cf. 6.3.2.1). Whereas grazing is described by O'Connor as search for some object which is marked by the clearly articulated query, browsing is the search without some object, therefore, there exists no way to articulate its character (ibid, p. 214). The second type of searching is especially conspicuous with activities related to creation of new knowledge (ibid, p. 222).

In the pragmatics of creating new knowledge according to O'Connor: "Because of the intent is to generate a new combination, there is no way to segment the collection. Short of engaging each and every document, a random sampling is made on the assumption that any location is just likely as another to yield fruitful results" (ibid).

The so called random searching in IR could perhaps only account for a small proportion of all searches, however its significance for the present discussion lies in another point rather than the question of its existence and use in IR.

One of the main arguments of sections 5.5.5 and 5.6.2.3 was that use-value of documents or rather of knowledge may not always be determined at the time of engaging with it, i.e. here and now. This amounts to saying that not all knowledge one acquires in day-to-day activities of one's being in the world *are put* in some positive use. Therefore, it should be made clear that one does not only engage with (read, recognize, etc.) documents for some positive value. It is also true that any knowledge that one acquires during one's existence contributes to his/her horizon (cf. 5.6.2.3, footnote 28), thus in some indirect manner eventually to one's relation to texts that one encounters in her/his professional and/or daily activities. However, this relation may not be obvious, or indeed evident. Certainly, not in terms of isolated linguistic entities such as words, phrases and such.

One might turn to phenomenology of Heidegger and Gadamer to explicate this process better (Winograd & Flores, 1986). The philosophy of Heidegger and Gadamer argues that our cognition of the world, as well as any hermeneutic process is a result of "Being" in the world ("Dasein"), i.e, a result of the condition of being "thrown" in the world, or "thrownness". Our understanding of the world is a result of our all previous interaction with the world, in other words previous experience of "Being". This constitute our "horizon" or "pre-understanding", in simpler terms "prejudices", which cannot at all made all explicit or rational. Interaction with texts is a hermeneutic process which is a result of being "thrown" in the world. Our engagement with texts find some resonance when we act upon them in the light of our horizon and the context of the moment of interpretation. Therefore, our relation to texts is a function of both our condition of thrownness and the horizon that we bring in, in interpreting them.

It is thus plausible to say that, all our existence, both professional and everyday is marked with our condition of being thrown in the world, which results in constant engagement in acting within a situation, some of which can be considered as semiotic in that they cause signifying processes.

The phenomenologic description makes clear that we not only search for documents and knowledge but we are exposed to them, thrown in them, in all sorts of ways even if we do not intend to. This is more obvious if we consider that, text should be understood as not only verbal texts, but all sorts of signifying practices as noted in section 4.2.3. Furthermore our engagement with texts takes place within the horizon we bring in. In other words, we interact with texts and interpret or do things with them by virtue of our horizon or pre-understanding. Therefore, our engagement with texts are made possible by our previous engagements. This was referred as intertextuality in 4.2.3.

The same cause-effect relationship applies to IR situation, as well. We come to IRS to search for more documents precisely because of our previous engagement with some documents, i.e. our pre-understanding. Only after some connection is made between documents one has already engaged with, that is, the relationship between them is prescribed (cf. 5.5.2, 5.5.5, 5.6.2.3), one approaches the system to describe (denote) more of like them. Therefore, the prescription of the relation, thus the relevance criteria is anterior to description and interrogation of a database.

It is therefore plausible to say that, in IR, as noted in 5.6.1.1, one starts from the text to formulate a problem and not vice versa. As detailed in section 5.6, searching begins when the user formulates a semantic model by picking-up the pertinent features from the continuous mass of textual data previously engaged with. Therefore, the initial prescription of what is wanted is an integral part of the search process. As pointed out in 5.5.5, the IR research for most of its part has only concerned with the part of the retrieval process after which the pertinent relations are prescribed. The equally important part of the IR process, in which the prescription of the relevance criteria is made has not been explicitly investigated. One of the reasons for this might be the implicit assumption that the relevance criteria are objective part of our knowledge of the domain, and not a construction (invention).

The so called random searching mentioned above can be seen in terms of the positivistic conceptualization of the scientific (research) pragmatics: There is a document somewhere in the database, when *discovered* provides the 'missing-link' for new knowledge. In O'Connor's words: "... the standard formal systems for representing documents often do not present to the researcher adequate means for discovering catalytic works. Browsing is a means for accomplishing such a discovery" (O'Connor, 1993, p. 212).

This may well be the case with some research activities, however arguments of the semiotic view of IR makes one to consider that, it may well be that inventive research is more to do with making something new out of what is one already exposed to (engaged with), making new connections out of what is already known (cf. 5.5.5, 5.6.2.3). In this regard, as it is with the phenomenologic account of everyday life, it might be that, creative labour in research pragmatics is more to do with being thrown in certain type of literature rather than others and creative ways of acting upon them (which implies a *methodological* approach rather than a random one in the final analysis; cf. Hjørland, 1997). This seems to be especially the case in the paralogical pragmatics of science (5.5.4, 5.5.5).

If the above account of creative labour is accepted, then browsing, i.e. willingness/openness to get oneself acquainted with new documents can be understand not as a random search for discovery but as a purposeful engagement that lasts until the researcher is able to invent a new connection, i.e. her horizon moves such that a new connection becomes apparent. The relevance judgement is then related not to the discovery of pertinent features but invention/prescription of new connections which may sometimes be tentative or even can finally be rejected by the fellow researchers in the field (cf. 5.5.5). In this respect, the inventive labour, far from being a trial and error or random search procedure, is a methodological labour well circumscribed by the concepts, tools, activities (in short, practice) of a given discipline (ibid).

6.2.2 Language games in Okapi

Okapi is a non-Boolean, interactive, information retrieval system based on a probabilistic model (cf. 7.2) which makes use of information derived from user's relevance judgements. The technical details of the system is given in 7.2.2. The logs of a typical Okapi session is presented in appendix R. In this section the course of interaction in a typical Okapi session is first described (6.2.2.1), then analyzed (6.2.2.2) in terms of the semiotic model of the IR process of chapters 4 and 5.

6.2.2.1 Interaction in Okapi

User starts off the interaction with Okapi by typing in some keywords describing her/his query. In the light of the discussion of the PRP (Robertson, 1977a) in relation to the retrieval model (Robertson & Sparck Jones, 1976) that Okapi is based on (cf. 6.1.1), query can be understood in terms of the concept of information need (Robertson & Belkin, 1978a). However, this interpretation is not necessary for the functioning of the retrieval model (cf. 6.2.2.2) and indeed this interpretation gave way to a more flexible one in its implementation in Okapi[8].

In a more recent conceptualization of the IR process in connection with the Okapi system, the search statement is understood as the description of the features of the documents that the user would like to see. This is evident in the phrasing of the request[9] for relevance judgement by the system when the user engages with one of the documents in the list produced by the system, which reads as : "*is this the sort of thing you are looking for?*" (cf. appendix Q).

After the user goes through the ranked list[10] and views the full record of some[11] of the documents, making a relevance judgement in response to the prompt "whether or not it is a sort of document s/he would like to see"[12] after each document viewed, the user has the option of initiating the automatic query expansion (AQE) facility of the system by choosing the "more" command (7.2.2).

AQE facility in Okapi works by taking the terms from the documents marked as relevant by the user and doing a new search based on these terms which may include the original user input search terms (for technicalities of these see 7.2.2). The assumption being that, the higher the odds of a feature appearing in a document marked as relevant, the higher the odds that it will appear in relevant documents that have not been yet seen by the user.

In a similar fashion, the user could examine some of the documents in the new list, making relevance judgements on each of them in turn. The user could then ask for more like the ones s/he has chosen as described above.

6.2.2.2 Analysis of the user-system interaction in Okapi

The objective of this section is to interpret the basic user-system interaction in Okapi, described in the preceding section in terms of the semiotic model of the retrieval process developed in this dissertation especially from the perspective of language games (cf. 3.9; 5.5.5).

The initial stage of entering the search terms as noted in 6.2.2.1 describes the features of the documents that the user would like to see. This is not different from saying: "(I would like to see) document(s) that contains following features (terms)" (in a boolean system the specification normally includes the spatial (topological) arrangements of the terms) (cf. 4.2.3; 4.3.2.3). In terms of the categories of language games, this is a denotative statement (cf. 5.5.1). It can be

[8] A recent discussion with S.E. Robertson during the writing of this dissertation made this clear. See the below paragraph for a description of this conception.

[9] Since the user must comply with this prompt by saying "yes" or "no", this should be considered as a command rather than a request (cf. 7.2.2).

[10] Produced as a result of best match retrieval with the terms taken from the user's search statement (cf. 7.2.2).

[11] At least one.

[12] Dichotomous yes/no type of judgement (cf. 7.2.2).

modulated into a request by prefixing it by a phrase such as put in brackets above, or: "Would you find the documents ...". This level of the interaction is called denotation$_1$ in 5.6.1 and the processes involved at this stage are analyzed in 4.3.2, 4.5.2, 5.4.1 and 5.6.1.

The next stage in the user-system dialogue in Okapi starts when the system searches the database and returns to the user those documents with the features described by the search statement. This is the second level of interaction referred to as connotation in 5.6.2.

It is then expected that, the user will scan the ranked list and select one from the list to examine it (read the abstract, so on; cf. 7.2.2) in more detail. When the user wants to go back to the ranked list, s/he will be prompted for a relevance judgement as described in the previous section.

The prompt is modulated as "*is this the sort of thing you are looking for?*" with the understanding that the system will find more of the documents like the ones the user has chosen by answering "yes" to the above question. This is therefore equivalent to saying to the user that: "I will find more of those documents that have common features with the document that you said 'is the sort of thing that I am looking for'". This in its turn implies that the system and the user have a common understanding of what constitutes the relevant features of the document(s) chosen and indeed the user would like to see more documents with those features.

Relevance feedback is therefore, denotation (description) of the features of the sort of documents that the user wants to see more of. This sort of dialogue is referred as denotation$_2$ in 5.6.2.1. The obvious reason for calling this level of interaction as denotation is that, the user is still in the descriptive mode, describing the sort of documents s/he would like to see. It is fundamentally a descriptive, hence, denotative labour. It should be clearly demarcated from the prescriptive labour (cf. 5.5.2; 5.5.5; 5.6.2.3) that will be discussed later in this section.

It is now necessary to discuss the relevance feedback stage in Okapi in more detail to explicate the characteristics of this language game. The first point to be examined is the sorts of possible responses that the user could adopt when prompted for a relevance judgement.

The relevance feedback prompt asks the user whether or not s/he would like to see more of the sort of documents s/he has chosen as described in the preceding paragraphs. The axiomatic of this prompt as noted above is that, the user and the system have agreed on or have the common understanding of what constitute the relevant features of the documents chosen by the user. Otherwise, it would not be possible to refer to the same object by the both parties. In other words, unless the system and the user both refer to the same features of the documents chosen by the user, it would not be possible for the system to come up with documents similar to those marked as relevant by the user. Since in Okapi, the system deals only with isolated linguistic entities such as words and phrases (cf. 7.3.2; 7.3.3) one is effectively referring to such features of the documents.

The most straightforward response one could have when prompted for a relevance judgement is either a "yes" or a "no". If the user does not want to see a document similar to the one the prompt is referring to, the answer should be "no". One should answer "yes" to the same prompt, if it is the sort of document that s/he would like to see more of, on the condition that, the user is referring to some isolated linguistic entities in the document, assuming that the documents with similar features have a good chance to be of some use, or more accurately to be the sort of documents that the user would like to see (see recognition at denotation$_2$ in 5.6.2.1).

However, there could be cases where none of the above two responses would apply. A simple such situation arises when the document one is referring to is a relevant document, i.e. the sort of document that the user initially wanted to find, but now having got it, the user does not want to see any more like it. In this case, the game of relevance feedback comes to a termination. Assuming that the rest of the documents in the ranked list is like the one that the user does not want to see any more of, and it is the design objective of the underlying retrieval model to produce a list just like that (cf. 7.2.2)[13], there would be no language game left for the user to perform other than quitting the system and starting off with a new search statement.

Another such case arises, when the user thinks that the document is relevant or want to see more documents like it, however, cannot say which feature(s) of it is relevant or significant for her. Obviously, this situation could only arise when the isolated linguistic entities are not (alone) useful/sufficient to discriminate the document from the others for the user. It is highly objectionable from the semiotic perspective of this dissertation that the meaning of a text is located in isolated entities (cf. intertextuality in 4.2.3 and Peirce's semiotics in 3.5.2, see also Thomas, 1993; and Blair, 1992; 1990). Thus, it can be legitimately expected (apart from the cases of pure didactics perhaps, cf. 5.6.2.1) that, the links that relate a document to the others similar in some way may not be easily specifiable in terms of isolated linguistic entities, therefore, an acute user who is familiar with system's retrieval model may find it impossible to give an answer to the relevance request.

More importantly perhaps, the user may find the document potentially useful but does not know in what way s/he can make use of it at present, i.e. here and now (cf. 5.5.5; 5.6.2.3). In the extreme case of the paralogical activities, any document could be of use, therefore, it would might not make sense to answer the relevance feedback request for each individual document (cf. 5.5.4; 5.5.5).

In all of the above cases, the language game proposed by the system (i.e. relevance feedback) is inappropriate and does not assist the user with the language game (speech act) that the user wants to proceed with, therefore could lead in some cases to the termination of the game.

In the first case described above for instance, where the user does not want to see any more documents like the system already found, yet has not been completely satisfied with the results so far, i.e. would like to continue with the interaction, there seems to be no obvious way for the system to proceed with if the user cannot define another query in terms of keywords/concepts (cf. O'Connor, 1993)[14].

One of the possible games that could be of use to the user in the above case is the so called

[13] Actually, the retrieval engine of Okapi ranks the documents according to decreasing probability of relevance to the user, therefore, the documents at the bottom of the list would be less likely to be similar to the one described by the original search statement, thus the user might do well in this case to look at the documents near the bottom. However, this would be of course the negation of the rationale of the ranked list (PRP).

[14] There could be two reasons for this; either the user's information need cannot be defined in terms of concepts (cf. O'Connor, 1993), such as in some instances of creation of new knowledge (cf. 5.6.2.3) or the user is unfamiliar with the domain and/or the vocabulary of the domain as in the case of the didactics situation (cf. 5.6.2.2). In both case, what is needed is perhaps to find a way to engage with some document, and act upon it in some (non-random) way (cf. 6.2.1.4), so that one can proceed with inventing (the former case) or discovering (the latter case) connections.

random sampling game as discussed in (O'Connor, 1993). However, from the point of view of this dissertation, this should be interpreted in terms of the discussion of 6.2.1.4, i.e. not as a random process as such, but in relation to our condition of thrownness, in other words, "Being" in the world, which implies methodological and conceptual framework of the discipline (knowledge domain) one is practising. In paralogical activities in particular, and in creative labour in general, one is actively looking for new connections, new interpretation of the known texts, that is, prescription of new relevance criteria[15]. To perform such a labour, one needs documents that s/he can interpret adequately (cf. the researcher, the expert, in 5.5.5 and 5.6.2.3), i.e. can act upon within her horizon (cf. 6.2.1.4). In the absence of such documents, the best labour one can perform is to find a way to systematically engage with some documents with a willingness to move one's horizon actively . This is browsing in opposition to grazing in the sense articulated by O'Connor (1993) (cf. 6.2.1.4). When one's horizon moves by engaging with some[16] texts such that s/he can prescribe new connections, the paralogical or the inventive, or the prescriptive game succeeds.

One can then perhaps go back to the descriptive game of retrieval ('prediction'). The conventional IRS normally support this sort of denotative labour. Okapi in its present implementation (cf. 7.3) supports only such a labour. Only after some relevant documents identified, i.e. the relevance criteria are prescribed (by the user), a probabilistic system like Okapi, or any system based on prediction, can find other similar documents (of course, once relevant documents are found, a competent user could perhaps do what a prediction or denotation based IRS does herself as well).

Similarly, in the other two cases described above, where there is no appropriate answer to the relevance feedback prompt, the denotative game of retrieval may not be the most suitable one for the user to continue with. Assuming that the user is trying to establish new connections (i.e. inventing, prescribing new relevance criteria) as hypothesized in the previous paragraphs, it should be useful to provide the user with tools that assist in the prescriptive labour. The question to be addressed then is not "how to get more information about the user's needs" (both the cognitive and probabilistic approaches articulates the problem in these terms; cf. 5.6.2.2) which is basically a question pertinent to a denotative game (cf. 5.5.2), but "how to aid the user in prescribing new links".

6.2.3 Discussion of the design practice as viewed from the semiotic perspective

In this section, following the articulation of the documentary information retrieval situation as a semiotic process in the preceding pages, the IRS design practice is discussed in terms of the semiotic model of IR developed in chapters 4 and 5.

The starting point for the present discussion of the design practice is to view IRS design as a

[15] 'Looking for it' should not be understood in terms of recognition of relevant features, but construction, i.e. prescription or invention of the relevant features as discussed in 6.2.1.4.

[16] The assumption here is that, there is no *a priori* set of such documents (cf. 5.5.5; 5.6.2.3). The best metaphor to describe the search situation where there is no pre-defined set of data is perhaps that of the condition of thrownness where one acts on the newly engaged texts with the intention of moving one's horizon (cf. 6.2.1.4).

social practice, i.e. as a practice taking place against social, political, and economic backgrounds, thus, should be analyzed/interpreted within such a context (cf. 4.3.2). Social practices are discursive formations that embody social relations (Frohmann, 1994; Poster, 1984). In this sense, they should be considered as fundamentally political structures. As any social practice, IRS design practice is therefore a discursive practice. This has already been noted in 5.5.5 and 5.6.2.3.

The IRS design practice in this regard should be understood as explicit decisions concerning implementation of particular language games or speech acts with a clear understanding of the economic, politic, and cultural purposes they serve in particular social contexts (cf. 5.5.3).

Although one can think of several different types of language games in an IRS, denotation and prescription are identified as the two most pertinent ones in terms of explicating the essential characteristics of the IR problematic (cf. 5.5). From the analysis of chapter 5, it appears that the main design decision is concerned with two contradictory system design practices: that of information transfer and knowledge production (cf. 5.6.2.3).

Information transfer is understood as transmission of already known (i.e. public) knowledge from a source such as a database or expert to a subject who is positioned as the addressee in didactics, i.e. student or trainee (cf. 5.6.2.1).

Knowledge production, on the other hand, concerns with establishment of new (not publicly known) connections between texts. In this regard, it is an inventive process in contrast the former's discovery oriented labour.

Furthermore, the above distinction highlights another important characteristic of the documentary information retrieval situation: whether the connections between texts are given (i.e. pre-known) as in the case of didactics, or invented as in the case of creative labour, it should first be prescribed. This has been noted clearly in 6.2.1.3. It has also been remarked in the same section and subsequently in 6.2.2.2 that most present IRS systems are designed to facilitate interaction after such a relation has been prescribed. The problematic of the process of prescription of the relevance criteria in the first place remains open to be addressed (cf, 6.2.1.4; 6.2.2.2).

It is the argument of this dissertation that, for most of its part, current IRS design practices take the relevance criteria as a given of the IR process with the accompanying implicit assumption that IR is solely about transmission of information or didactics (cf. 5.6.2.2; 6.1.2; 6.1.3).

From the above ongoing discussion, IRS design practice emerges as "decisions concerning the choice between transmission vs. production of knowledge or denotative vs. prescriptive language games".

In conclusion, system design practice as viewed from the semiotic perspective of the present dissertation emerges as decisions regarding the implementation of different types of language games in relation to the two basic labour of denotation and prescription (transfer of information and production of knowledge). In this connection, it might be useful to summarize the denotation/prescription or transmission/production axis of the IR labour.

Transmission of information relies on the denotation (description) of the information need of the user. Without a description of the information need, information transfer cannot be actualized. However, before the user's information need can be described, it should be first prescribed by an anterior labour. This labour could take one of two forms; it is either assigned to the user (didactics case) or invented (production of knowledge case).

Production of knowledge is a process which involves the prescription (invention) of the relevance criteria by the user. Contrary to the case of the transfer of information, in this labour information need (relevant features of the documents) is prescribed by the user (i.e. not assigned to the user[17]).

The above formulation of the design pragmatics will guide the discussion of the design objectives of the knowledge-based systems in the next section (6.3).

6.3 Discussion of the Design objectives

The aim of this section is to formulate the main design objectives of the knowledge-based systems (KBSs) described in chapter 7 by applying the semiotic view of the documentary information retrieval situation developed in chapters 4, 5 and 6 to the givens or constraints of the design environment, i.e. the given tools (Okapi IRS, cf. 7.2; Inspec thesaurus, 7.1) in a given context (the institution it is used in; the user population, 8.2.2.3).

However, before proceeding with the formulation another design constraint should be introduced. Due to the practical limitations of available time and resources, it is acknowledged that the programming needed for system development has to be kept to a minimal level and instead of a highly desirable full interactive interaction with the system, limited interactivity through batch processing is to be implemented.

In section 6.3.1 the operational context in which the systems designed are discussed in terms of the semiotic view of the IR interaction. This is followed by the formulation of the main design objective of the implemented systems in section 6.3.2.

6.3.1 Discussion of the operational contexts

The sort of user interaction available in Okapi IRS is discussed in 6.2.2.2. The analysis of the same section reveals the types of language games implemented in the system.

To summarize the discussion of 6.2.2.2, it was noted that, there are two modes of interaction with Okapi, i.e. two language games: *a)* the initial formulation of the search statement by keywords (cf. 5.6.1.1) *b)* the following stage of relevance feedback, which involves recognition of documents as pertinent to the particular information need (cf. 5.6.2.1). Both of these games are belong to the generic class of games called denotations (cf. 6.2.2.2).

It is remarked in 6.2.2.2, it is possible to think situations in which neither of the above games would be appropriate for the user to perform. One of the simpler cases where such a situation

[17] As a subject of didactics, the user must learn to denote correctly the relevant features of the documents required in terms of the indexing and retrieval rules of the system. The above statement is especially valid for the cases of strictly 'subject access'. Many factors normally influence the users' relevance judgements even in the case of didactics or subject access (see footnote 7). Therefore, even in this case relevance criteria is partly prescribed by the user, however this point is of secondary importance for the purposes of the discussion here.

arises is when the user enters a search statement and is presented with a list. After examining some of the documents in the list, the user does not want to see any more like them. As Okapi could only find documents like the ones chosen by the user from the list, it cannot offer the user any other game to perform.

It is hypothesised in the same section that in the above situation, having found enough number of relevant documents, the user does not want to "see any more like them", however, her information need has not been resolved by the documents found so far. The user therefore, wants to see some other sorts of documents, which she is not able to specify in terms of keywords/concepts.

It is argued in 6.2.2.2 that, the reason for not being able to formulate a new search statement could be due to one (or both) of the following : either the user is unfamiliar with the subject domain, therefore, not familiar with the domain's vocabulary and concepts, or engaged in an inventive labour therefore, no *pre-specification* of the relevant features is possible. In both cases what is needed by the user is presented with a tool which would enable her to formulate new relationships between documents or some attributes of documents, so that s/he can act upon them (cf. 6.2.1.4).

To address the above described problem, it is first necessary to identify the type of the game needed to resolve the situation, i.e. is this a denotative or a prescriptive type of game? In other words, can the documents/attributes required by the user be predicted? The answer to this question depends on the type of labour: if the user engaged in a creative labour, it could be hypothesised that, the sort of documents s/he wants to see next may not be predicted, except of negative specification such as, "not the class(es) of documents that the user is already familiar with" (O'Connor, 1993, p. 222). If, on the other hand, it is a didactics situation, it could be hypothesised that, the relation between the user's initial search statement and what s/he requires now, after seeing some documents from the original list may be inferred (cf. 5.6.1.2). The above hypothesis regarding the relation between the original user query and the subsequent information need is dealt with in the next section.

In the rest of this section, the hypothesis concerning the relation between the above described problem case and the operational context in which the new systems tested, is discussed.

The user population of Okapi is mainly consists of MSc and undergraduate students from the School of Informatics, although substantial number of research students from the school and other schools and departments of the university also use the system. In addition to this, academic members from several departments also use Okapi. The user population thus constitutes quite a heterogenous group (see also 8.2.2.3).

It is reasonable therefore to assume that, the users of Okapi approach the system with a mixed bag of information needs, ranging from queries resulting from the assigned course work for the MSc and the undergraduate students, to vague or unspecified needs as a result of creative research activity for some others.

In terms of the evaluation of the new systems with real users with real needs drawn from the above described population, one should make assumption(s) about the nature of the expected queries. It is difficult to expect from the size of the active users of Okapi at any given time and the duration of the evaluation exercise to have reasonable number of user's with information needs exclusively about inventing (prescribing) new connections. It seems much more reasonable to assume that both the MSc/undergraduate and the researcher population approach to the system with a mix case of information needs, involving some pre-specified (i.e. prescribed) relevance criteria as well as some hope to come across something not foreseen initially (not prescribed), i.e. with willingness to explore the topic of interest in a larger context than originally formulated and foreseen in the search statement (cf. 6.2.1.4).

It is reasonable to say that this willingness to explore an area larger than the original search topic could be a result of, either the user's involvement in a creative labour, thus, the desire to move her/his horizon (cf. 6.2.2.2), or the original information need being multi-faceted, consisting of partially overlapping queries (fig. 6.3).

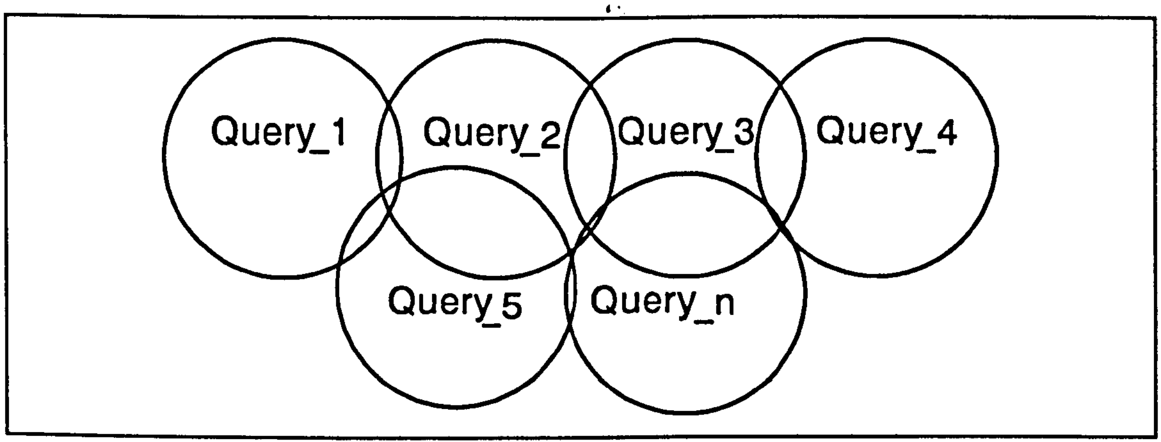


Figure 6.3: Partially Overlapping Queries

In both of the cases, it is likely that the user having found some documents corresponding to his initial query would like to find some other sorts of documents (related to the original query in obvious or non-obvious ways) which he could not specify due to reasons described in the beginning of this section. Out of all possible cases discussed in 6.2.2.2, this seems to be the most likely information game that the users of Okapi would like/need to perform.

6.3.2 The overall design objective

The most important characteristic of the game described above, whether as a result of assignment or willingness to explore is that, there are more than one prescribed relevance criteria applicable during the interaction with the system. In other words, the user has a set of partially overlapping (see fig. 6.3) descriptions of what sorts of documents s/he would like to see. Each of these descriptions *prescribe* an individual relevance criteria which may/may not overlap with all the others, furthermore, which may or may not explicitly known to the user at the time of approaching to the system (i.e. some of which may emerge as a result of the search process).

If the above formulation of the retrieval situation is accepted, then in a probabilistic retrieval mechanism (cf. 6.1.1), the objective of the system may be formulated as "to optimize the system's performance for each of the individual relevance criteria separately".

In the wake of the PRP, which constitutes the basis of the retrieval model of Okapi (6.1.1; 7.3.4), I would like to *re-formulate* the above as: "*documents should be ranked according to the probability of being pertinent to the knowledge domain defined by the user in separate categories of relevance (i.e. individually for each part of the knowledge domain), in parallel (i.e. synchronically)*". This, I will call the 'Probability Ranking Principle in Parallel' or 'PRPP'.

Number of things in the above definition need clarification. It is first of all hypothesised that, the user approaches to the system with some query that is related to a knowledge domain. However, the boundaries of this domain is entirely up to the user (cf. 5.6.2.3). It is also hypothesised that the knowledge domain can be divided conveniently into non-exclusive sub-domains which constitute parts of the whole domain (or its 'facets'). This is in accord with the encyclopedic model of the universe of knowledge presented in 5.7.2. Each facet of the general domain does not have a fixed boundary, but defined by the discourse (produced by the research) in the domain, therefore changes (moves). The structure that underlie the production of discourse in a given domain is modelled in 4.3.2. One of the pragmatics of science, namely research, is actually for its substantial part about the definition of this boundary (cf. 5.5.5).

The individual relevance criteria from which the document descriptions derive represent a particular sub-domain (or some parts of it) in a given general knowledge space.

The objective of the system is to rank the documents according to their probability of being relevant to the user within each individual relevance category separately. Assuming that the relevance categories are pre-defined by the user, then the system should present the documents from each relevance category to the user in the order of her choice, or at the same time (i.e. no particular order, in parallel). In the absence of such user defined categories, it might be desirable to rank the relevance categories according to their probability of being conceptually close to the original query statement (cf. 7.5.4).

In terms of the particular language game to be implemented in Okapi (cf. 7.5), the above translates as: "*from the initial search terms produced by the user, suggest (indefinite number of) alternative terms grouped together such that, represent some part of the knowledge domain related to the domain defined by the user's original terms. Optimize the ranking of documents in each relevance category (or for each group of terms). Rank these alternative groups of terms (batches) in the order of their probability of being related to the domain defined by the user's original search terms*".

The most important point in the above definition is that, it moves the goal of the IRS from predicting the terms that are most closely related to the original search term(s), to suggestion of a '*family of alternative groups of terms*' that are related to the user's original query to varying degrees. It is true that we still want groups of terms that are related to the original terms, however, the plural form of the word group above implies clearly that what is required is more than just *the* best alternative group of terms or the best batch. It prescribes that those that are less likely related to the original terms should also be made available. In other words, not only those known (or inferred) to be related to the original query should be brought to the attention of the user, but also those that are less likely related. This is why the above formulation worded as 'suggestion' of the alternative group*s* of terms, rather than prediction of the most likely group of terms. The subtle difference between the two is that, in suggesting one is actually *prescribing* the relevance criteria, rather than merely predicting what has been *a priori* prescribed. This prescription is certainly based on the relatedness of the subject(s) described by the user input terms and the suggested groups of terms (or batches), however, it should also be possible to specify the way the user input terms are related to the suggested groups of terms (batches) so that the user could divert the search from finding 'documents similar to those initially described by the search terms', to 'those very or completely dissimilar'. Once the path is cleared from the obstacles that limits the horizon of IR exclusively to the denotative labour, i.e. description (therefore, implicated prediction) of "what one likes to see", one can go to the extreme case of searching for totally (previously) unrelated domains, i.e. the case of radical inventions (cf. 5.7.2).

The justification for this approach rests on the assumption that the user at the time of approaching the system initially or after seeing some documents is interested in seeing documents not only 'like' as initially conceived and anticipated, but also documents from domains closely or loosely related to the one originally described in the search statement (cf. 6.3.3.1). This could be either due to the user's engagement in a creative labour as discussed in 5.6.2.3 (i.e. prescription of new relations between documents), thus, *willingness* to *take in* more than it is necessary *now* or, due to the nature of the query (i.e. being multi-faceted from its inception, cf. 6.3.1).

It seems like the only way to show the user the terms that are loosely connected with the original query as well as those are more closely associated with it, is to group them in small numbers which as a whole makes sense to the user, i.e. *distinguishable* from one another by the user in some way (cf. 7.5). This would enable the user to abduce the contents of the database before actually seeing the documents in the database[18]. The *indefinite number* of ways (within the limits allowed by the systems' index language; cf. Bookstein, 1989) of doing this (i.e. generation of indefinite number of batches, cf. 7.5.2) is necessary if we do not want to limit ourselves exclusively to the cases where the relevance criteria are *pre-defined* or *known*[19] (cf. 6.3.3.1).

As it is assumed in 6.3.3.1, users of Okapi have mix bag of information needs, it is important to keep the alternatives as open as possible. The ranking of the groups of terms (batches) according to their probability of closeness to the original search statement should take care of users' with limited scope of information needs.

The significance of the above formulation for the actual systems designed (7.5) needs to be briefly noted here. It prescribes, regarding the selection of the documents from the database for retrieval, the inclusion of those less likely to be related to the original description of the documents wanted, on the assumption that, those not described by the original query might still be of (potential) use to the user, or more accurately, the user might also like to see such documents. The justification of this assumption is discussed in the preceding paragraphs and in 6.3.1. The KBSs of chapter 7 do this, as it will be shown in 7.5.3, by selecting from a knowledge-base (cf. 7.1) not only those terms that are conceptually nearest to the user input search terms, but also those related by some conceptual distance. This is done by producing

[18] The adoption of the thesaurus-based approach in the design of the KBS (cf. 7.1) makes sense in this connection. Rather than showing the user the documents themselves, some attributes of them are shown, hoping that this enables the user to judge whether the documents with those attributes may be the sort of thing that they want to see. An additional benefit of using a thesaurus is that, it records conceptual relations in a knowledge domain systematically. Therefore, it is a suitable device for inventive labor which is, as noted in 6.2.1.4 and 6.2.2.2, not a random but a methodological process.

[19] There is nothing preventing us in theory to enable (or suggest to) the user to go to areas totally unrelated to what is described by the original search statement (as noted earlier), as long as there is a way for the user to find her way around the database. This is where presentation of attributes of the documents rather than themselves should prove to be most helpful (cf. footnote 18).

separate batches which as a whole makes sense to the user in cognitive terms[20] (i.e. as describing a certain subject or number of related subjects; see 7.5.3). It is the assumption of the design that, the document list produced by an individual batch would then reflect the same cognitive coherence of the batch itself as describing a certain subject or number of related subjects (see 8.2 and 8.3).

6.4 Overview of the Systems Design and Evaluation Objectives

In this section, first the design objectives of the systems developed in this project (which are described in chapter 7) are reviewed (6.4.1 and 6.4.2). Subsequently, the objectives of the experiments performed to evaluate the systems (chapter 8) are explicated (6.4.3).

6.4.1 Semiotic theory and systems design

To explicate the systems design objectives, it is first necessary to summarize the semiotic theory developed in the preceding pages which has guided the systems design and evaluation parts of the project.

Very briefly, semiotic theory developed in this project argues that, interaction in IR can be characterised by two distinct class of activities: a) knowledge/information transfer or transmission of knowledge/information, b) knowledge production or creation of new knowledge (see especially chapter 5 and sections 6.2.1.3, 6.2.1.4).

In the first of the above cases, the user has a full or partial prescription of the required features of the documents searched or in other words the *relevance criteria*. There could be various reasons for the user not having full prescription of the relevance criteria. The user may not be familiar with the knowledge domain (its contents, and properties such as vocabulary and terminology etc.). On a more trivial level, the user may not be familiar with the commands and structure of the retrieval mechanism. All this may cause the user to have difficulties in formulating the relevance criteria and thus denoting the documents that satisfy these criteria. However, in this case whatever may be the reasons for the inability of the user in denoting the correct documents, it can be assumed that the relevant documents can be objectively (or inter-subjectively; e.g. panels of experts doing the relevance judgements) identified (cf. chapter 5; see also Karamuftuoglu, 1996; 1997).

There may be however several factors that influence the user's decision in accepting some of the documents identified by the experts in the domain as relevant, and rejecting others as not-relevant (cf. Karamuftuoglu, in press). Therefore, subjectivity of the user inevitably interferes with the objective (inter-subjective) assessments of the experts in the domain. These factors include the structure and the style of the document, presentational details (such as graphic or

[20] This is a very important requirement of the semiotic view of IR. User's understanding/cognition of the batches is part of the code which constitute a semiotic/linguistic system not totally dissimilar to natural languages. The functionality of the retrieval systems conceived in this project support more than the goal of finding just the documents that the user wants. See the discussion in 6.2.1.1 in this regard.

tabular summary of the results v. textual summary etc.), inclusion of a particular data or formulae, currency of the data included, and many others. While the above mainly cognitive factors may not disqualify a particular document's membership to a particular subject domain, they may be crucially important for the individual user in solving a particular problem in an actual situation.

In the second of the above cases (i.e. knowledge production) there is a totally different problem. In this case, the user most likely possesses all necessary knowledge regarding the domain. This should be a valid presumption, as production of new knowledge requires full competence in a particular subject domain. The difficulty in this case arises from the fact that, since this is an activity of production of new knowledge, by definition the documents necessary for this task are not *a priori* known by anyone, including the fellow experts in the domain. Otherwise this would not be production of new knowledge. Since which documents will satisfy the user are not known beforehand, the user does not have the relevance criteria before interacting with the system (except maybe the prescription of what sort of documents are not wanted). Therefore, some sort of interaction with the retrieval mechanism other than keyword/concept-searching is needed to enable the user to discover documents to help with the inventive labour involved in the knowledge production process. This sort of approach to interaction is generally subsumed under the umbrella term of browsing.

It is argued in chapter 6 and in particular in sections 6.1.2, 6.1.3, and 6.2.2.2 that, most of the cognitive and statistical/probabilistic approaches to interaction, in particular the PRP which constitutes the underlying philosophy of the Okapi system used in this project, are not very helpful in the above described situation of knowledge production. The reason for this, such systems limit the search to the user prescribed area of the document collection (or the knowledge space). In other words, such systems generally rely on the initial determination and denotation of the search space and try to optimise the search within this space which is explicitly or implicitly defined by the user.

Some other systems do not require such a pre-determination and fixing of the search space. 'Social information filtering systems' or as sometimes called 'recommendation systems' for instance do not rely heavily on the user specified search domains, but make use of other social and contextual cues (see Karamuftuoglu, in press). Maltz & Ehrlich (1995) for instance report a system which enables colleagues to share newly found documents by sending them directly to those who might find them interesting. The system relies on the existence of group of people who know each others' interests and willingly distribute and share new information. This system, therefore, does not rely as much on the pre-determined search area defined by the user as the existence of an inter-subjective bonding among a group of people who discuss potentially interesting new information and suggest newly found documents to their colleagues for consideration/interpretation.

The crucial point here is that, the search area in such systems is not limited to a domain determined by a single user, but dynamically and collectively (inter-subjectively) negotiated and discussed, therefore, sudden and unexpected jumps or leaps in the subject area searched can happen. Such unexpected diversions from the original thread (search subject) usually cannot be foreseen by any individual member of the community alone but emerge as a result of complex interaction patterns (discussion, negotiation, persuasion etc.) among the participants.

The systems designed in the present project aim to provide the users with retrieval mechanisms which, similar to social filtering systems, enable them to explore new areas that may be partially or totally unknown to them previously in the framework of the Okapi system. The users may find some of the suggested search areas useful and some not. Furthermore, the relevance of the some of the suggested areas to the users' search area (in the context of the systems designed in this project as implied by the user's original search terms; cf. 6.3.1 and 6.3.2) may not be obvious, i.e. experts in the field may not conjecture a connection between the user's search domain and some of the system suggested search areas. This however does not preclude the potential usefulness of such search domains. It is argued in 5.6.2.3 and 6.2.1.3 that, positing of such new connections is the distinctive feature of the inventive labour. Some other system suggested areas may however be highly obvious choices for an expert in the domain. This may not be so for the enquirer, however, for the reasons mentioned previously.

The above arguments put the general objectives of the systems designed in the present project in the framework of the semiotic analysis of the retrieval interaction developed in the preceding pages. In the sections that follow below, how this general objective is implemented in the particular context of the Okapi probabilistic retrieval system used in the project will be discussed.

6.4.2 The design objectives

As discussed in 6.3.1, Okapi can only help the user if the user can denote the search space by some keywords or identify some relevant documents from the list of documents retrieved by the system. However, as argued in the above paragraphs and in 6.3.1, this is useful in finding documents in a partially or fully prescribed search area. Actually, Okapi is particularly useful if the user cannot define the search area fully but have a partial or general prescription. The initial prescription of the search area is important, as the system relies on this evidence in the first place to conduct a search in the document collection. In a knowledge production situation it may however be necessary to conduct searches in areas outside what has been prescribed by the user, as argued earlier. As knowledge production labour is an inventive act, what documents (therefore the search spaces) will be useful in the final analysis, are not normally known to the user or indeed anyone else.

As discussed in 6.3.2, in the context and limitations of the present project, one way of suggesting new areas of potential interest to the user is to present groups of conceptually related terms that describe contiguous subject areas (see 7.3 and especially 7.3.2). The purpose of presenting to the user a number of related subject areas instead of a single subject area that is most closely related to the user's original search area (as defined by the search terms) is to do with the underlying system design objective in this project which aims to enable the users to explore new domains.

To achieve the above stated objective, a thesaurus (Inspec) and the relationships between the terms embedded in the thesaurus are used to construct a knowledge-base or a semantic network representing the subjects contained by the documents in the document collection and the relationships between them. By starting from the user input terms, the knowledge-based systems (KBSs) designed generate clusters or as called in this project batches of conceptually linked terms in the thesaurus that are as a whole (as a batch) related to the user input search terms. As users usually utilise a few search terms (see 7.3.1.4), the original search space is often very general. It is therefore likely that by expanding the users original query by adding a number of linked thesaurus terms the user's query becomes better defined or more specific (cf. 7.3.2; 7.3.3; 7.4.3).

The batches are generated in the KBSs by finding two thesaurus terms that are best candidates for representing two distinct aspects implied by the user's initial search terms (7.3.1). These terms are referred to as the 'source terms'. The next step is to connect the two source terms (7.3.2) by following the various relationships that are available between the terms in the Inspec thesaurus (7.1.1). This results in a batch of linked terms consists of the two source terms and a number of intervening terms or 'nodes' connecting the two source terms.

An important design decision that differs from other systems which use the *semantic network* approach is that, a number of redundant nodes (terms) are introduced in the path connecting the source terms as necessary (7.3.2.1 and 7.3.2.2), to increase the likelihood of finding documents that are not foreseen by the user's original search terms. Most systems using the semantic network approach in information retrieval (cf. 2.2.2, 7.3.2.2, 7.4.2) aim to find in a thesaurus the most closely related terms to the user's original query. For the reasons discussed above, in this project, redundant terms are purposefully added to divert the original search to other areas that may be useful to the user. Since in the semantic network approach to IR, it is generally assumed that (2.2.2), the distance (i.e. number of intervening terms) by which a thesaurus term separated from a given term is indicative of the conceptual closeness of the two terms, by including terms that are separated by a longer distance (higher number of nodes) from the source terms to the batches generated, likelihood of finding documents not foreseen by the original user search terms are increased. The extra terms added to the batches are not necessary to connect the two source terms and are referred in this dissertation as the 'redundant terms'.

Introduction of the redundant terms also result in more specific batches that define finer portions of the knowledge space (7.3.2.1, 7.3.2.2). This makes possible to perceive better the structure of the knowledge space defined by the original search terms, as well as, to help conceptualize different aspects that original search terms may have. It may be appropriate to compare the individual batches to maps of particular areas of a more general space defined by the user's original search terms which border each other and collectively map the whole of the terrain. This is hypothesised to aid the user to understand the structure of the general search space and help explore aspects of the search space that may not be known to the user previously (7.3.2.1). This aspect of the batches can be useful both in a knowledge production task, where the user explores new and previously unrelated domains, as well as in a knowledge transmission task, where the user learns about the structure (the concepts and the relationships between the concepts) of a domain which s/he may not be familiar previously.

It should also be noted that, it is assumed that by showing related terms in a batch, any ambiguity a term might have when shown on its own to the user is removed. In semiotics, as discussed in 3.5.5 and 4.2, value of a term in a sign system derives from its position relative to other similar terms (see also 7.3.2). This semiotic principle supports the above indicated assumption regarding the usefulness of the batches in removing potential ambiguities associated with single terms and was effective in deciding to deal with batches of related terms in this project rather than single thesaural terms as in the CILKS project (cf. 7.3.1.2 and 8.4.2) for instance.

The evidence from one of he CILKS experiments suggests that, users often make mistakes in judging the contents of thesaurus terms when shown on their own (see 8.4.2). This is especially a problem in a large knowledge domain covered for instance by the Inspec thesaurus, where similar terms are used in totally different contexts. An indirect evidence that supports this observation comes from the responses of the users' to the questions related to hypothesis no. 3 below (see H3 in 6.4.3) in the evaluation experiments performed as part of this project. The results of the experiments indicate that (see 8.3.1 and 8.3.2), some of the users participating in the evaluation experiments perceived some of the batches as representing new ideas to them although they did not choose any of the terms constituting those batches as representing new ideas when shown on their own.

6.4.3 Hypotheses, Questions and the Evaluation Experiments

To achieve the above broadly outlined objectives, two systems referred to as *KBS-1* and *KBS-2* are designed (chapter 7) and two experiments, experiment 1 and experiment 2, are performed to evaluate these systems (chapter 8).

Both *KBS-1* and *KBS-2* aim to present to the users number of batches of contiguous terms that represent knowledge domains that are related to the user's original search terms. The main objective of the experiments which are designed to evaluate KBSs is to find out whether the users consider the batches useful in representing their original query and whether any of the batches represent new ideas to them (i.e. initially the user did not think of). A related goal is to find about the effectiveness of the batches in retrieving relevant documents.

A number of hypotheses are generated in this connection:

1. The first hypothesis need to be tested is whether the batches generated by the system perceived as representing distinct subject areas by the users. More specifically, whether the users perceive different batches as representing different subjects (although they must be related to each other somehow as they share at least the two source terms).

To evaluate this aspect of the system, the users participated in the experiment are asked to assign one or more of the eight categories to the top ranking ten batches generated by *KBS-1*. The assumption here is that, if the user applies different categories to differentiate between the batches in the list, it can then be concluded that the user has detected sensible differences between them. The categories provided include, "the batch as a whole looks good for my users search purpose", "some of the terms in the batch are good for my search purpose", "the batch contains terms that represent new ideas which are useful to my search", "the batch contains term(s) that represent ideas which is/are part of the general domain of my search, however not directly useful to me", "none of the terms in the batch are good for my search purpose" (see appendix A for full details, also see 8.2.2.1).

Therefore, the first hypothesis to be tested is:

H1: Batches represent distinct subjects that can be detected by the users

The evidence to evaluate this hypothesis comes from the users judgements on the batches using the eight categories mentioned above. This aspect of the evaluation exercise is discussed in 8.3.1.2 and 8.3.2.2 (see also 8.3.1.1, 8.3.1.3, and 8.3.2.1, 8.3.2.3).

2. The second important hypothesis to be tested about the batches is whether batches represent well the users' original query and whether they help initiating new ideas.

The second hypothesis is therefore:

H2: Batches are useful in helping the users to define better their query and/or stimulating them in following new conceptual relations to explore new search domains

As a batch that represents the user's original query (search area/subject) well and a batch that represents new ideas to the user cannot be the same batches (i.e. these two categories are logically exclusive of each other), this hypothesis implies that users *may* have more than one distinct queries, i.e. more than one relevance categories (see the discussion of PRPP in 6.3.2). The users' judgements on the batches using the categories mentioned earlier provides the evidence to test this hypothesis. The results for this part of the evaluation exercise are discussed in 8.3.1.2 , 8.3.1.3 and 8.3.2.2, 8.3.2.3.

3. Thirdly, it would be useful to know how useful the batches as a whole in stimulating the users to explore new ideas in comparison to single thesaurus terms.

The third hypothesis, therefore, can be worded as follows:

H3: Batches are better at suggesting new ideas to the users compared to single thesaurus terms

For this purpose, individual unique terms extracted from the top ranking ten batches are presented to the users (prior to the stage where the batches themselves are shown) in alphabetical order in a single list, and asked to indicate any terms that suggest new ideas to them. The users' responds for this part of the experiment are compared with their assessments of the batches. The results for this part of the experiments are discussed in 8.3.1.2 and 8.3.2.2.

4. The second stage in both experiment 1 and 2 aims to find out about the retrieval effectiveness of the batches. The hypothesis tested in this stage of has three related parts.

4-a. The first part of the fourth hypothesis is about the effectiveness of the batches in clustering different relevant documents. It has already been hypothesised that batches represent distinct subjects (H1) and users may have more than one distinct (although related) search subjects or queries (H2). From these two hypotheses it can therefore be expected that, different batches cluster different relevant documents. The first part of the fourth hypothesis can, thus, be worded as follows:

H4-a: Different batches retrieve different relevant documents (different relevant documents are clustered in different batches)

The evidence to test this hypothesis comes from the user's relevance judgements on the documents retrieved by the batches, and the distribution of the relative positions of the relevant documents among the batches used in the experiments. The results for this part of the experiments are presented in 8.3.1.4 and 8.3.2.4. The relative positions of the relevant documents in the lists produced by the batches provide the necessary information to test the above hypothesis.

4-b. A related hypothesis to the above is that, relevant documents are clustered in some of the batches and not all of the batches used in the search process. It has been hypothesised that batches represent distinct subjects (H1) some of which may be categorized by the users as non-relevant (as indicated by the category descriptions such as, "the batch contains term(s) that represent ideas which is/are part of the general domain of my search, however not directly useful to me", "none of the terms in the batch are good for my search purpose"). It can therefore be expected that some of the batches are likely to cluster documents that are not of interest to the users. In other words, it can be expected that, some batches cluster the non-relevant documents while some others the relevant ones. Therefore, we can formulate the second part of the fourth hypothesis as follows:

H4-b: Some batches do not cluster the relevant documents

The evidence to test this hypothesis is provided by the users' relevance judgements and distribution of the relevant documents in the document lists generated by the batches. The results for this part of the experiments are presented in 8.3.1.4 and 8.3.2.4.

4-c. Given hypothesis 4-b which states that some batches cluster the non-relevant documents, it can be expected that, users can distinguish between the batches that cluster the non-relevant documents from those that cluster the relevant documents by looking at the contents (i.e. the terms included in the batches and their conceptual relations) of the batches. In other words, it can be hypothesised that, users can recognize/select the batches that contain the relevant documents and reject the others by looking at the terms included/excluded in each batch. The third part of the fourth hypothesis is worded as follows:

H4-c: Users can select (recognize) the batches that contain relevant documents and reject the others that do not (relevant documents are clustered in the batches indicated by the user as containing terms that are good for the user's search purpose).

As it has been hypothesised in H1, batches represent distinct subjects which can be detected by the users. Users express these differences by putting batches into different categories (using one or more of the eight categories provided for each batch). Some of these categories assert that a given batch as a whole is, or some of the terms in a batch are useful (relevant) to the user's query or represent new (useful/relevant) ideas, while some other categories assert the opposite, i.e. a given batch is not useful (non-relevant) to the user's query (cf. wording of the eight categories in appendix A). The users' useful/not-useful (or relevant/not-relevant) judgements on the batches using one or more of the eight categories mentioned above and the actual relevance judgements done by the users on the documents retrieved by the batches, provide the evidence necessary to evaluate this hypothesis. The results for this aspect of the evaluation exercise are discussed in 8.3.1.2, 8.3.1.4 and 8.3.2.2, 8.3.2.4 (see also 8.3.1.3, and 8.3.2.3).

***KBS-1* and experiment 1**

The above hypothesis from **H1** to **H4** are common to both systems and therefore to both experiments performed. However, there is one additional hypothesis investigated in experiment 1 alone.

5. It is hypothesised that, batches are better in clustering relevant documents than a set of individual thesaural terms that does not bring together related terms. For this purpose, retrieval effectiveness of the individual batches are compared with the effectiveness of the set of terms formed by taking individual terms from the four top ranking batches used in the first experiment. The hypothesis is formulated as follows:

H5: Single batches are more effective than a list of individual thesaurus terms in clustering the relevant documents

The results for this part of the experiment are presented in 8.3.1.4.

6. Different formulae can be used to rank the generated batches. It would be useful to get some indication of the relative ranking abilities of different formulae. Therefore, a further question is formulated to find out whether F4 formula used in *KBS-1* (see 7.3.2.3) is better than the other candidates.

Q1. Does F4 have a better ranking ability than other formulae used in *experiment 1*?

The results for this part of the experiment are discussed in 8.3.1.4.

KBS-2 and experiment 2

KBS-2, similar to *KBS-1*, aims to present to the user number of batches of contiguous terms that represent knowledge domains that are related to the user's original search terms. Therefore, the main objective of experiment 2 which is designed to evaluate *KBS-2* is similarly: to evaluate the usefulness of the batches in representing the users' queries and initiating new ideas, as well as their retrieval effectiveness.

7. There is only one significant difference between *KBS-1* and *KBS-2*. *KBS-2* aims to have a high retrieval effectiveness that is comparable to that of Okapi which is taken as a benchmark system (see 7.3.2.3). There is therefore one important difference between the objectives of experiment 1 and experiment 2. In experiment 2, performance of the individual batches are compared with Okapi's performance. In addition to hypotheses H1 to H4 above which are tested in both experiments, a further question is tested in experiment 2 which is not addressed in experiment 1:

Q2: Do batches have a level of retrieval effectiveness better than Okapi?

Note that in experiment 2 retrieval effectiveness of the batches generated by the system is compared with the Okapi system, whereas in experiment 1 it is compared with a list of individual thesaurus terms. In 8.3.2.4 results for this part of the experiment are analyzed.

Hypothesis H5 and question Q1 in experiment 1 are tested to aid the design of *KBS-2*. The results of these tests are taken into account (in choosing the ranking formula for the batches, and deciding whether to use batches or a list of thesaurus terms derived by selecting unique terms from individual batches) in *KBS-2* (see 8.3.1.4; see also 7.3.1.3, 7.3.2.3). Therefore, in experiment 2 the above noted hypothesis and question are not reexamined.

Chapter 7
Knowledge-Based Systems For IR

In the following sections, knowledge based systems (KBS) designed and implemented as part of this project are described and discussed in detail. In section 7.1, Inspec thesaurus as a knowledge base is discussed. Section 7.2 presents the details of the Okapi retrieval system. In section 7.3 various components of the designed systems are described in detail. Section 7.4 presents a general discussion of the various parameter and decisions involved in the systems designed. An overview of the design objectives and hypothesis of the KBS can be found in section 6.4.

7.1 The Knowledge-Base

A thesaurus is a device to record relationships between terms that represent concepts in a knowledge domain to facilitate vocabulary control in indexing and searching (Jones, 1993).

It has been known for a long time that terms assigned by different indexers or by the same indexer at different times vary to a great extent (Paice, 1991). This is commonly known as inter-indexer inconsistency (Borko, 1964; Zunde & Dexter, 1969). The same is true at the searcher end of the retrieval process. Users have been observed to use greatly varying vocabulary in denoting desired documents or representing their information needs (Saracevic & Kantor, 1988; Furnas et al., 1987).

It has been therefore generally accepted that, the main difficulty in document retrieval is the selection of correct terms for denoting concepts in indexing and searching. The traditional response to overcome this problem has been to attempt to control the use of vocabulary in a domain by compiling a standard list of terms authorized for use in indexing and retrieval (Paice, 1991; Jones, 1993)

Such an authorized vocabulary often recorded in artifacts known as thesauri which prescribe the terms to be used for indexing and searching and the relationships between them in a knowledge domain or a subject area.

In the following section first, description of one such artifact, the Inspec thesaurus is presented (7.1.1). This is followed by the description of the computer stored version of the Inspec thesaurus in 7.1.2. Finally in 7.1.3, use of the Inspec database as a knowledge-base is discussed. The information presented about the Inspec thesaurus and its computerised form held in Oracle relational database draw mainly from the internal report produced for the CILKS project (Jones, 1992).

7.1.1 The Inspec thesaurus

There are three main term relationships in the Inspec thesaurus as in most other thesauri. These relationships are: equivalence, hierarchical and associative.

The *equivalence relationship* exists between a *preferred* term and a set of *lead-in* terms. Lead-in terms are not actually used in indexing documents, therefore do not enter into the other two types of relationships, the hierarchical and associative, with the terms in the thesaurus. They are used for pointing the terms, called the preferred terms, that are actually used for indexing documents and which may enter hierarchical and associative relationships with the other preferred terms making up the thesaurus.

The equivalence relationship is construed to gather synonyms that refer to the same or closely related meanings. According to ISO standard 2788 (in Aitchison & Gilchrist, 1987), the equivalence relationship covers variant spellings, abbreviations, acronyms, popular forms of scientific terms, and so on.

Ideally, the equivalence relationship should be of the type one to many, i.e. a single preferred term equated with a number of lead-in terms. However in the Inspec thesaurus this principal does not hold true. The Inspec thesaurus permits many to many equivalence relations. There are 800 lead-in terms in the Inspec thesaurus which have more than one preferred term. Twenty-five lead-in terms have five or more preferred terms and one has thirteen. The preferred term(s) for a lead-in term is indicated by the USE marker in the Inspec thesaurus. Conversely, a lead-in term is designated by the mark UF (Use For).

It has been noted that (Jones, 1992), the relationship between some preferred terms and lead-in terms in the Inspec thesaurus can hardly said to be that of synonymy. Although this can be a problem for some automated thesaurus navigation systems such as the one developed in the CILKS project (Jones, 1993; Jones, et al 1995), it does not constitute a specific problem for the knowledge based approach developed in this project for the reasons explained in section 7.3 below.

There are in total 12787 terms in the Inspec thesaurus, of which 6191 are preferred and 6596 lead-ins. There are 23661 term relationships of the types, equivalence, hierarchical and associative between these terms (Jones, 1992).

Most of the Inspec terms are *compound terms* rather than single words. In total 77% of all Inspec terms are compounds (Jones, 1993). Thesaurus construction manuals recommend use of single words instead of compound terms (Aitchison & Gilchrist, 1987), however, in practice most terms used in thesauri are compounds. This is partly to do with the fact that single words are highly ambiguous in isolation. Compound terms are needed for indexing precision (Jones, 1993).

Another device to increase indexing precision is *term qualifiers*. Term qualifiers are given in parenthesis next to the thesaurus terms to distinguish between different senses of a term or delimit the precise field of reference of an already specific term. The latter seems to be the predominant use of the term qualifiers in the Inspec thesaurus (ibid). However, this device is not widely used in Inspec. There are just 456 qualifiers in the Inspec thesaurus, only 65 of which are for preferred terms (ibid).

The *hierarchical relationship* establishes genus-species affiliation between terms. Narrow terms (NT) are listed below their broader terms (BT) forming a tree structure. However as noted in (Jones, 1992), Inspec hierarchies do not conform to strict tree format. The reason for this is that, a narrow term may have more than one parent in the Inspec thesaurus, thus, belong to more than one hierarchy (up to six as noted in Jones, 1992).

A common way of forming a narrow term from a broader one is to prefix the broader term with another qualifying adjective (sometimes called a 'difference word'). Alternatively, the previous adjectival phrase may be replaced by one with a more specific meaning (Jones, 1993). As one goes up a term hierarchy, the coverage of the terms become more inclusive (general). According to the above cited ISO standard, the hierarchical relationships may involve, genus-species, whole and part, category and instance, or geographical inclusiveness (in Jones, 1993).

There are 537 hierarchies in the Inspec thesaurus, ranging in size from two (one broader and a one narrower term) to 213, with an average of 12.5. Number of levels in the Inspec hierarchies range from 2 to 8, with an average of 2.9. The total number of hierarchical links (relationships) in the Inspec thesaurus is 5906.

Many hierarchies in the Inspec database are connected at the lower levels as a result of multiple parentage of many of the narrower terms. The inspec thesaurus is therefore fully connected through hierarchical links with the exception of perhaps a few isolated pockets (Jones, 1992).

The *associative relationship* connects two terms that are conceptually related. This relationship can be used to identify such associations as between a thing and its application, an affect and a cause, an activity and an agent of that activity, a thing and its parts, and so on (Lancaster, 1986, 46-47). The associative relationship is reciprocal and is indicated by the RT mark.

The associative relationship is semantically looser than that of the hierarchical. There are almost twice as many associative links in the Inspec thesaurus (10155) than the hierarchical ones. Number of terms linked by the associative relationship increases exponentially with the number of levels expanded from a given term. On average, each term in the Inspec thesaurus has 3.7 (maximum 35) immediately related terms (i.e. at one level of expansion), this increases to 22.6 (maximum 266) at two levels of expansion.

It is most likely that, the associate links connect the Inspec thesaurus completely, i.e. every term in the Inspec thesaurus is connected with every other through the associative relationships (Jones, 1992).

The Inspec thesaurus classifies the preferred terms, in addition to hierarchical (broader-narrow) term relationships described above, according to a set of class-codes.

Most Inspec preferred terms are assigned one or more class-codes which produce a hierarchical classification independent of the broader-narrow term relations. There are 2420 class-codes in Inspec arranged in four main classes (Jones, 1992):

A - Physics
B - Electrical Engineering & Electronics
C - Computers & Control
D - Information Technology for Managers

Many of the preferred terms have more than one class-codes. Majority of them are A/B only. The document collection used in this project (see 8.2.2.4) and the CILKS project comprise of the documents from the section of the Inspec database with class-codes C or D only.

However, these subdivisions are not self-contained with respect to term relationships discussed earlier (ibid). When the preferred terms with class-codes A or B (4716 of them) are removed from the thesaurus, 1475 terms remain. Of these 238 have immediate links to A or B classified terms through one of the three term relationships discussed earlier. About 30 of them have links only to those with A or B class-codes. More importantly perhaps, many of the documents in the document collection used in the project indexed by the terms with A/B class-codes.

7.1.2 The data-model of the thesaurus database

Oracle relational database is used to store the information embodied in the printed Inspec thesaurus. The Inspec thesaurus database is developed for the CILKS project. It has been adapted and used in the present project.

Logical data model of the thesaurus database is given in Jones (1993). It is reproduced below:

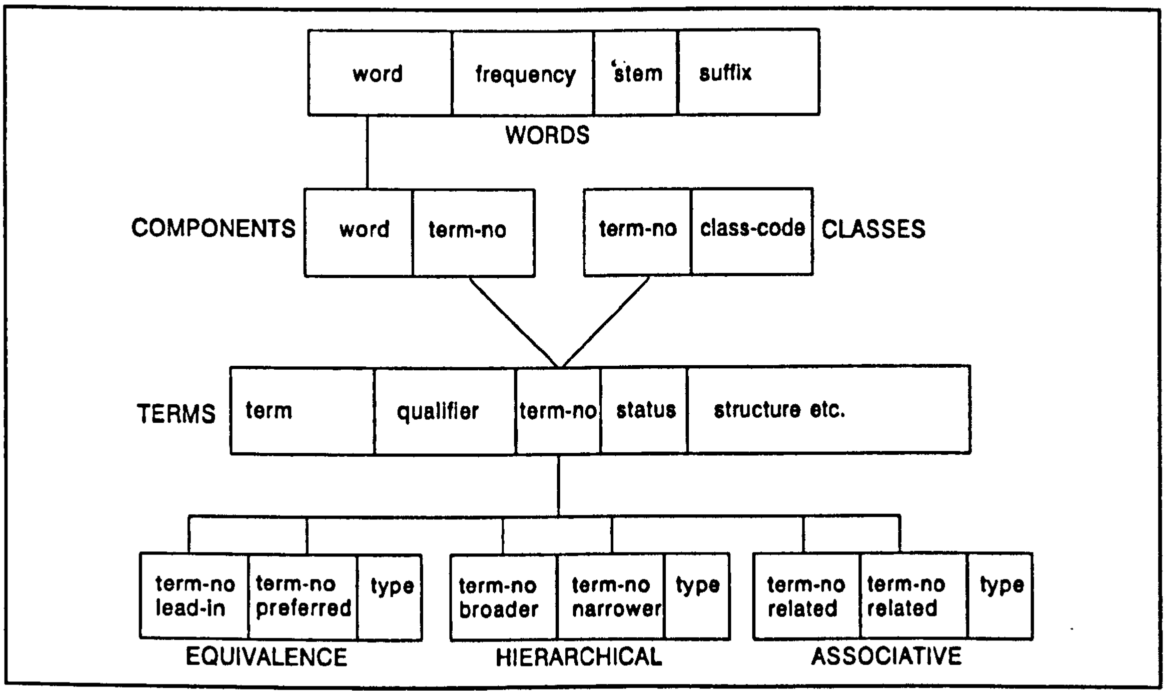


Figure 7.1: Logical Data Model for the Thesaurus Database

The exact text of each term as it appears in the printed form (including any capitals, hyphens, apostrophes, etc.) is stored in the table TERM (figure 1). Each term is stored once in the database and identified by a unique number (Jones, 1992; 1993).

The TERM table contains the following information for each term (Jones, 1992; 1993): qualifier, status, structure indicator, number of documents indexed by the term, and the term number.

Qualifiers are kept separately from the main text of the term. Inspec thesaurus makes limited use of the term qualifier. There are only 456 of them in Inspec, and in 70 cases they distinguish between otherwise identical terms (Jones, 1992).

The *status* field holds information regarding the relationships the terms enter with the other terms. The basic division here is between preferred and lead-in terms. Preferred terms are further sub-divided according to their position in the term hierarchy.

The Inspec thesaurus already has the concept of 'top terms'. Top terms are those without a broader term in the term hierarchy, i.e. the broadest or the most general term in the hierarchy. In the Inspec database preferred terms are coded according to their position in the term hierarchy as follows (ibid):

Top: these terms are with at least one narrower term but with no broader term. There are 537 of them in the Inspec thesaurus.

Middle: these are with both narrow and broader terms. Total number in the thesaurus is 1217.

Bottom: these terms do not have any narrow terms but have a broader term. There are 3570 bottom terms in the thesaurus.

Related: these are with neither narrow or broader preferred terms which are connected to the other terms only through the associative relationship (RT). There are 867 such terms in the thesaurus.

As noted in 7.1.1 most of the Inspec terms are compound terms. There are 4858 compound preferred and 5103 lead-in terms in the Inspec thesaurus. The *structure indicator* field holds information regarding whether or not a term is a compound term.

Number of documents indexed by each term in the document collection is also held in the TERM table.

Each term is represented by a unique key field, the term number, in the database which links together term information held in the different tables shown in figure 1 above.

Most preferred terms in the Inspec thesaurus are associated with one or more *class-codes* as discussed in 7.1.1. This information is held in the thesaurus database for each term in the table CLASSES (figure 1).

The database also holds information regarding the individual words that make up the terms in the Inspec thesaurus in the table WORDS. This table holds both the stemmed and unstemmed forms of the words and includes information regarding the frequency of occurrence of the words in the database and the compound terms in which they occur.

The information held in this table is used to identify partial matches between the user input free-text search terms and the thesaurus terms (see section 7.3.1 below).

To determine individual words in the Inspec thesaurus, heuristics are defined to identify word boundaries. In practice this amounts to decide whether two words separated by a hyphen should be treated as a single word or two separate words. Briefly, "a hyphen is treated as a single word-boundary unless the word on either side of it begins with a capital letter, is less than three letters long, or contains a digit or another non-alphabetical character such as a slash. Thus strings like "air-traffic", "alpha-particles" are treated as two words, while "add-on", "X-ray" and "Bose-Einstein" remains as one" (Jones, 1992, p. 4).

Total number of individual words are found to be 6561 (ignoring the two stop words, "of" and "and") in the database when the hyphens are removed following the above described heuristics. Word frequencies in the database range between 1 and 294 with an average of 4.5 (ibid, p. 5). Stemming of the individual words are done following the standard procedure used in Okapi (see 7.2.2). Each term is linked to corresponding entries in the WORD table through the COMPONENTS table (figure 1).

Finally, the database holds information regarding the equivalence, hierarchical and associative relationships for each term. These are held in three separate tables as illustrated in figure 1 above. Each relationship is held only once in the database (Jones, 1992). A more detailed discussion of these three types of term relations is given in 7.1.1.

7.1.3 The thesaurus as a knowledge-base (semantic net)

The idea of treating a thesaurus database as a semantic net for information retrieval purposes has been discussed in section 2.2.2. In this section some specific points related to the use of the Inspec database described in 7.1.2 as a semantic net in this project are addressed.

As noted in section 2.2.2, each term in a thesaurus database can be considered as a *node* in a semantic network. In semantic networks, a node is connected to others via 'links'. In a thesaurus database, the obvious candidates to substitute for the function of a link in semantic networks are the term relationships such as, equivalence, hierarchical and associative described in section 7.1.1 (see 2.2.2). However, there are a number of differences between a conventional semantic net and a thesaurus database as a semantic net that worth noting here.

The first main difference between a conventional semantic net and a thesaurus as a semantic net is the size or the breadth of the knowledge domain covered by the two. While a typical semantic net in AI applications deals with a narrow domain knowledge, a usual thesaurus maps the concepts in a broad domain, typically covering several related subject fields (Jones, 1993; cf. also 2.2.2). As noted in section 7.1.1, the Inspec thesaurus contains 12787 terms and 23661 relationships between these terms which cover the general knowledge domain of computer science and information systems & technology.

Another important difference between the two forms of semantic nets is that, in a conventional semantic network nodes and links have unequivocal structures which prescribe their precise uses. Whereas in a conventional thesaurus most of the information regarding the structure and use of the terms and the term relationships are not explicitly encoded in the artifact. As stated in Jones (1993), a thesaurus is a repository for a great deal of human knowledge and its compilation is based on logical and explicit principles, however much of this information is implicit in the thesaurus and it is not available for automated manipulation. Use of a conventional thesaurus indeed demands contextual interpretation of the terms and their relationships which is so difficult to replicate artificially (ibid).

Some semantic net applications used in AI and related fields assign a score to each individual link connecting two nodes to represent the strength of association between them. This information is then used to control the operations performed (e.g. as the spreading activation) in the network (cf. 2.2.2). In information retrieval, some of the systems that make use of semantic network type of structures employ this principle to measure the distance or strength of association between two terms. This information is then used to find terms close or similar to user input terms (Chen, 1992; Chen & Dhar, 1991; cf. 2.2.2).

Some of the IR systems, such as the Topic system (Chong, 1989) manually assigns such scores to the links, while others compute them automatically according to statistical properties of the knowledge-base and/or the type of the relationships (Chen et al., 1995; 1997; cf. 2.2.3). While manual assignment of scores by the user or the system developers might be feasible for a small knowledge-base covering a narrow domain, for a general thesaurus like Inspec this is clearly not a feasible option. Heuristics used for automatic computation of such scores although applicable for larger knowledge-bases, not used in this project, as the main underlying philosophy of the project aims to provide maximum flexibility in exploring different subject areas without becoming totally unstructured therefore useless. As noted in section 6.3.1, major design principle in this project is to avoid making too many assumptions about the user's information needs in order to enable them to expand their original search interests and explore new areas. In the retrieval systems designed in this project, groups of terms or batches carry far more importance than the individual terms that make up the batches (cf. 6.4.2). It is therefore uneconomic and perhaps unnecessary in the implemented KBSs to assign scores to measure the strength of association between the terms in the knowledge-base in a more precise way. It is important in this project not to overcode (overdetermine) the knowledge-base and the retrieval system in general. It is preferable in this study, to use the terminology of semiotics, to undercode (cf. 4.4.1) the system. In section 7.4 some related points to this issue are discussed in relation to the design parameters of the retrieval systems developed in this project.

7.2 The Okapi retrieval system: an overview

The Okapi system which is the in-house document retrieval system used for research in the department provided the platform to develop and test the knowledge-based systems in this project. There are two knowledge-based systems (KBSs) developed and evaluated in this project which are described in detail in section 7.3. In this section the Okapi system is introduced and its structure is outlined. In section 7.2.1 an overview of the Okapi system is presented. Section 7.2.2 outlines the probabilistic retrieval function used in Okapi.

7.2.1 Okapi Overview

Okapi is an interactive experimental document retrieval system based on a probabilistic retrieval model. Interaction with system is done via different layers of interfaces built on top of the Basic Search System (BSS) which provides the lowest level of protocols to access the system (Robertson, 1997).

The Okapi system is used in various projects within the department to develop and test new retrieval models and conduct research related to information seeking behaviour of users. The interactive version of the system which is operational continually since 1989 at City is used by registered users of the system who access it over the Campus Wide Network of the university.

The interactive system available over the university network is character-based (VT100) and does not support GUI. The registered users of the system which belong to different departments of the university (both student and staff) can search the City University library catalogue, the Inspec database which is used in this project, and the University of Bath library catalogue.

To monitor the use and access of Okapi databases, users are registered and issued user account names and passwords which are entered every time to log into Okapi. All transactions between the users and the system are logged and kept in files for each user accessing the system with their registered user ids. This information is used in investigating the user retrieval behaviour and the effectiveness of the Okapi system by the researchers affiliated with the department. An example log of a session is given in appendix R. The users took part in the evaluation exercises in this project are identified and selected using the transaction logging facility of the Okapi system (see section 8.2.2.3 for more on this point). Detailed information regarding the structure and functionality of Okapi can be found in Walker & Hancock-Beaulieu (1991), and Robertson (1997).

7.2.2 Okapi retrieval process

The retrieval process in the interactive Okapi system consists of following steps:

- The user input search terms are preprocessed, parsed and stemmed

- The search terms are assigned weights

- Corresponding document weights are calculated for the search terms

- a list of document surrogates (title, author, date) with decreasing weights is displayed to the user

- user may select a document from the ranked list to read the full record (abstract and related fields such as descriptor terms, source, etc.)

- when user wants to go back to the ranked list of retrieved documents, the system requests a binary relevance judgement

- the user may activate the automatic query expansion facility of the system after some relevant documents are found

- user's query terms are expanded according to information obtained from the documents selected as relevant by the user

The above outlined steps of a session with interactive Okapi are described in more detail in the following paragraphs.

Preprocessing, parsing and stemming of user input search terms

The user input terms are preprocessed to remove capitals, hyphens, punctuation and similar linguistic devices to convert the user input to the standard form used in indexing the documents in the databases. A user query such as "Hypertext and CD-ROM databases U.K." would become "hypertext and cdrom databases uk". After input terms are preprocessed they are parsed to remove stop words and identify common phrases stored in the go/see list (GSL).

The GSL performs some of the functions of a thesaurus, although information contained in this file is very scanty compared to conventional thesauri. It is mainly used for identifying synonymous terms such as "united Kingdom, uk, great britain, gb, britain", stop words such as "the, a, an", common prefixes, such as "pre, anti", and phrases that constitute single lexical units, such as "information retrieval, expert systems". Once the stop words are removed and words that match with the entries in the GSL are determined, the remaining user input words are stemmed. User input terms found in the GSL are treated separately according to their category in the list. Finally, the stemmed words and the GSL terms in the user input are looked up in the indexes and number of matching references for each index term are determined.

Weighting of terms and the probabilistic model

The probabilistic model of the Okapi system aims to predict the probability of a given document being relevant to the user's query by calculating weights for each document. The principle behind this method is known as the "probability ranking principle (PRP)" (Robertson, 1977a;

see also 6.1.1). Document scores are calculated from the sum of the weights of individual query terms indexing that document.

The probability of a document indexed by the term 't' being relevant to a given query is calculated by the following formula (F1):

F1: $w_t = \log \{ p_t (1-q_t) / q_t (1-p_t) \}$

where,

w: is the document weight indicative of the probability of the document being relevant to a query

p_t: is the probability that the term t will occur in relevant documents

q_t: is the probability of the term t will occur in non-relevant documents

The probabilities p and q are estimated as follows:

F2: $p_t = r/R$ and F3: $q_t = (n-r)/(N-R)$

where,

n: is the number of documents in the database (collection) containing the term t

N: is the number of documents in the database

R: is the number of documents chosen as relevant

r: is the number of relevant documents containing the term t

Substituting F2 and F3 to F1, we have F4 'selection value' of the term *t*; w_t:

F4: $w_t = \log\{(r+0.5)/(R-r+0.5)\}/\{(n-r+0.5)/(N-n-R+r+0.5)\}$

In the above equation 0.5 is added for each of the components of F4 in order to avoid indeterminate values and increase accuracy when there is little relevance information. When there is no relevance information, as it is the case at the beginning of a search session when the user enters new search terms, F4 reduces to:

F5: $w_t = \log\{ (N-n+0.5)/(n+0.5) \}$

The full derivation of the above formulae is given in Robertson & Sparck Jones (1976).

Ranking of the documents and relevance feedback

After the term weights are calculated using the above formulae document scores are calculated simply adding up the weights of the query terms that index it. Documents are presented to the user in descending order of their scores. Documents with same weights are ordered chronologically and within that in alphabetical author order.

The user are presented with brief details of the documents initially. This information consists of the title (or part of it), the author name and date of the document. Users can select a document from this list to see the full record of it. In the Inspec database, these include the abstract and the source of the document, as well as free-text keywords taken from the abstract and descriptor terms assigned by the indexers from the Inspec thesaurus.

Once the user reads the full record of a document, she/he is prompted to make a relevance judgement about the viewed document by the question "is this the sort of thing you are looking for?". The user must answer "yes" or "no" at this prompt. This relevance judgements are used in formula F4 above when automatic query expansion (AQE) of the system is activated. When the user chooses at least one document as relevant, AQE facility can be activated. The user can choose to activate this function by typing the character-key "M" ("Type M to see more books similar to the ones you have chosen"). AQE automatically extracts terms from the documents chosen as relevant by using the F4 formula. This function is used to modify the user's original query terms and produce a new ranked list of documents.

Terms in relevant documents are weighted by F4 (in the current implementation by WPQ, see below) and sorted in descending weight order. Those terms above a certain cut-off point are used to retrieve new documents. There is also a parameter that can be set to limit the number of terms to be included in the expanded query. The rationale behind the relevance feedback process and AQE is that terms found in relevant documents are also likely to be found in retrieving additional relevant documents.

It has been suggested by Robertson (1990) that in selecting terms from the relevant documents for AQE, the frequency of a term t in the relevant set should be taken into account to increase the retrieval effectiveness. The modified form of F4 which makes use of this information is known as 'WPQ selection value' and is used in calculating the term weights and document scores. WPQ is calculated by the following formula:

$a_t = w_t \, (p_t - q_t)$

Where, w_t, p_t and q_t are calculated as in F4, F2 and F3, respectively.

After AQE, document scores are calculated by the above formula, and the retrieved documents are ranked in decreasing order of their scores. Documents that are already viewed (both marked as relevant or not-relevant) in the previous iteration are removed from the new list and remaining documents are shown to the user.

7.3 The Knowledge-Based Component

In this section structure and design parameters of the 'knowledge-based component' (KC) are described and discussed (see figure 7.2 and 7.3).

The design objective of the KBS is laid down in 6.3 and 6.4. The overall goal of the new system is to provide the user with *alternative lists of documents* derived from different clusters (batches) of search terms, each batch representing some part of the knowledge space related to the space defined by the user's initial query terms. It is hypothesised that this would enable the user to cope with the situation where s/he does not want to see any more documents like those described by the initial search terms, however, does not know how to go about looking for different sorts of documents for the reasons discussed in 6.2.2.2 and 6.3.1.

In the context of the Okapi IRS, a possible implementation of the above objective is formulated and discussed in 6.3.2 and 6.4.3. According to this, the KC produces (*indefinite* number of) batches or clusters of terms, in which terms are conceptually related to each other so that each batch represents a part of the knowledge domain. More accurately, each batch makes a coherent set such that, it can be distinguished from others by the user conceptually/semantically. These batches are then used for searching the database, and produce ranked list of documents for each individual batch separately.

It is hypothesised in the same section that, conventional thesaurus offers the capability of presenting some of the attributes of documents in a document collection such that it makes sense conceptually to the users of the thesaurus and can be used as a clue to the contents of the collection. It is therefore assumed in this study that, batches of linked terms as described above can be produced satisfactorily by using a thesaurus which functions as a knowledge-base representing the knowledge space defined by the document collection, such as the one developed for the CILKS project (cf. 7.1). The overall structure of the resulting KBS is given in figure 7.2.

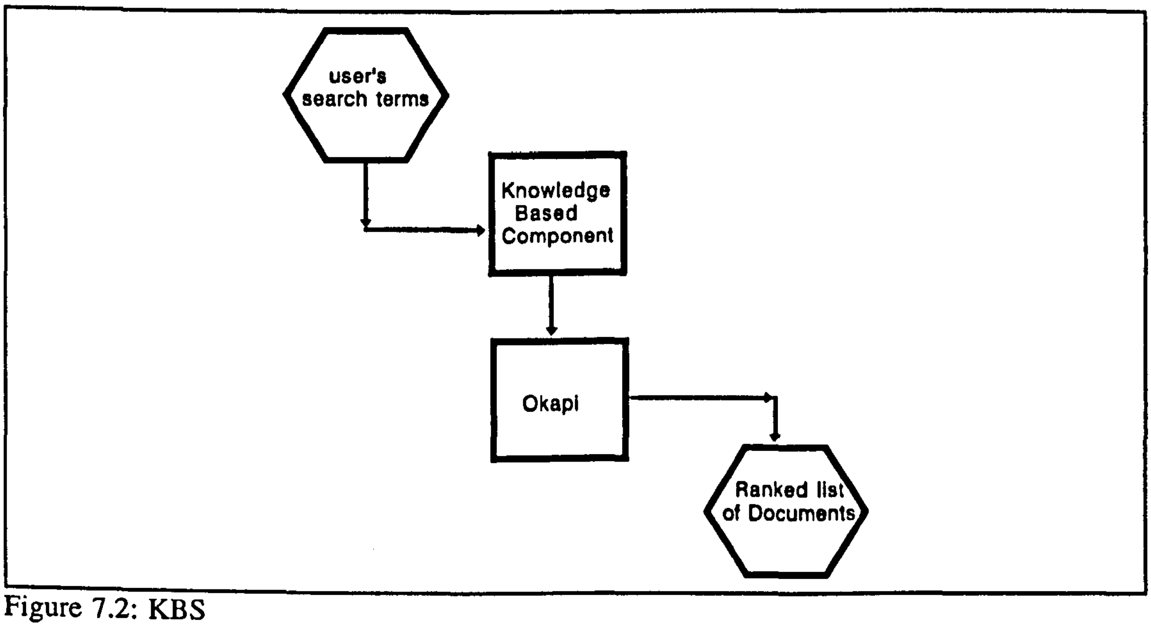


Figure 7.2: KBS

In the first stage of the KC (fig. 7.3), the user input terms are matched against the terms of the Inspec thesaurus, whose structure is discussed in 7.1. The procedure to select the source terms from the matched thesaurus terms is explained in 7.3.1. The mechanism of generating batches of linked terms from the source terms and ranking of the batches according to the probability of their relatedness to the user input search terms (cf. 6.4.3) are discussed in section 7.3.2. Section 7.3.3 describes the re-formulation of the user's search statement taking into account terms derived from the knowledge-base, the mechanism for searching the database, and production of the ranked list of documents.

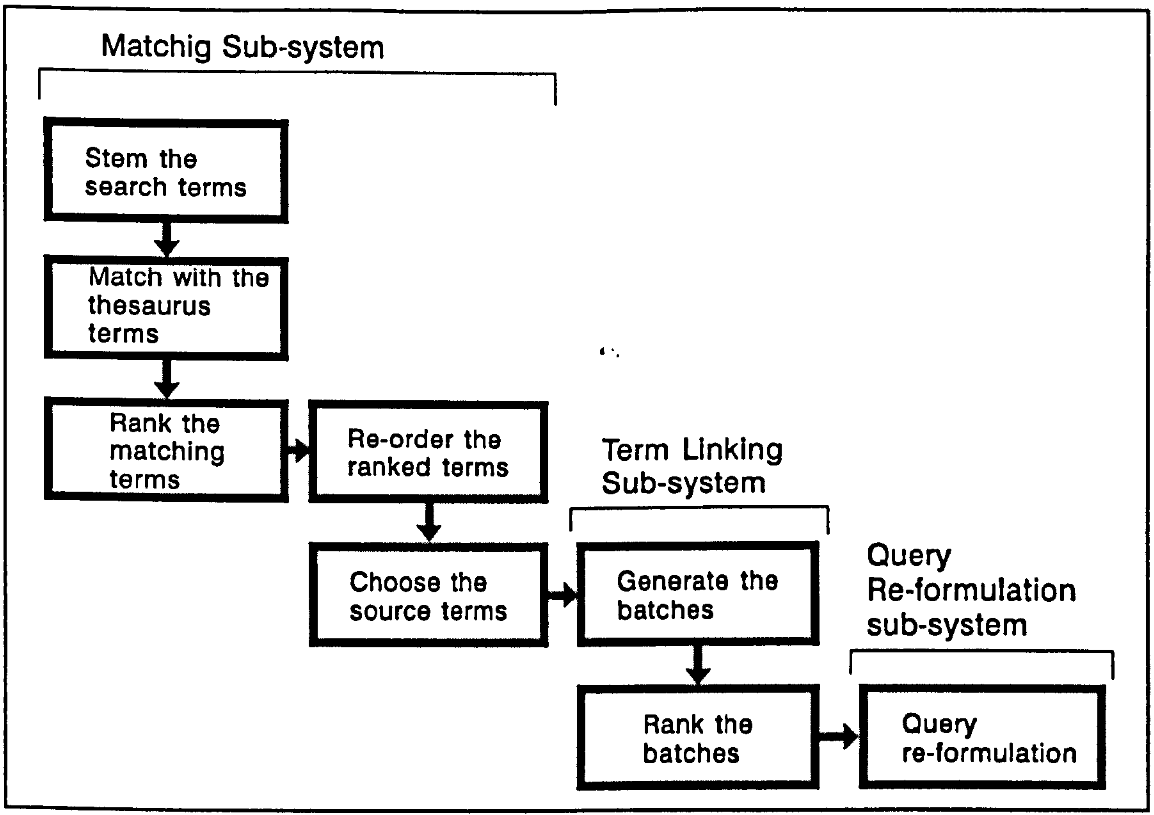


Figure 7.3: The Knowledge-Based Component (KC)

7.3.1 The source terms

The KC aims to offer the user different sets of search terms (batches) which are distinguishable from each other in cognitive terms, and describe a part of the document collection. One way of doing this as pointed out above is to compare the user input terms against the terms in the thesaurus, and select those which best represent the user's query. The basic assumption in this procedure being that, a rough equivalence can be established between natural language query terms and controlled vocabulary of a search/index device such as a thesaurus (Paice, 1991).

The procedure of selecting equivalent terms from a reservoir of controlled vocabulary such as a dictionary or a thesaurus is often referred as 'normalization' of the search terms (Smeaton, 1991; Paice, 1991; Vickery & Vickery, 1993). There are several methods of establishing equivalence of natural language and controlled terms. Some of these methods are examined in 2.2.1. The method used in this project is the subject of the present section (7.3.1).

The selection of the *source terms* in the KC is a two-step process (fig. 7.3). The first step involves matching the user input terms with the thesaurus terms (7.3.1.2). The subsequent step involves selection of the thesaurus terms that best represent the user's search terms (7.3.1.3 and 7.3.1.4).

7.3.1.1 Search term normalization

Various means of normalization of search terms have been surveyed in 2.2.1. The particularities of Okapi search environment however precludes straightforward adoption of any one method reported in the literature.

Most normalization techniques involve, either interaction with the user for clarification of the user input or designed bearing in mind specifically systems working with the Boolean logic (see 2.2.1 and 2.2.2). For the reasons explained in 6.3, the system designed in this project is not an interactive system, therefore, clarification of the normalization process by consulting the user is not possible.

Furthermore, Okapi is a best match system (cf. 7.2) which does not require formulation of the search statement by combining terms with logical operators as in Boolean systems. This results in highly unstructured search statements often involving elimination of duplicate terms at the time of input of the search terms by the user. Okapi does not even require the order of search terms conform with some of the syntactic or semantic rules of the natural language. This means that the user input in many instances cannot be adequately/correctly analyzed by standard natural language processing techniques (cf. 2.1.3 and 2.2.1).

The above mentioned peculiarities of the design environment made necessary to develop a mechanism of term normalization specific to the design setting of this project which is described in the following sections.

7.3.1.2 Matching with the thesaurus terms

The user input terms in Okapi often do not conform to syntactic and semantic rules of natural language texts. This makes it prohibitively difficult for syntactic analysis such as performed in systems like TomeSearcher (Vickery & Vickery, 1992), and IOTA (Chiaramella & Defude, 1987) (see also 2.1.3, 2.2.1). The difficulty arises, generally speaking, in delimiting the terms and group of terms that form phrases (e.g. noun phrases).

One simpler method of overcoming this as developed in the CILKS project (Jones, 1995; 1993; 1992) is to first match individual terms in the search statement with the terms in the thesaurus, and then rank the all matching thesaurus terms according to degree of match to the *whole* of the search terms. The degree of match takes into account two parameters: number of terms from the user's input matching the thesaurus term and the weight of the thesaurus term as determined by the idf values of its component terms within the thesaurus database. The steps involved in this procedure is described below:

A) Stem the search terms:

The user input search terms are stemmed as usual using the Okapi's routine for this (7.2.2).

B) Match with the thesaurus terms:

User input terms after stemming matched with the stems of the thesaurus terms held in the WORDS table of the thesaurus database (7.1.2).

C) Rank the matching terms *(Produce a ranked list of thesaurus terms matching the user input terms)*:

The matching thesaurus terms in step B above are ranked first, according to the number of user input terms (stems) they contain. The terms in the same rank group are then sorted by the ascending order of number of component terms they contain (i.e. if two terms have the same number of matching user terms, the term made up of fewer number of words is ranked above the term made up of more words). Finally, terms that are in the same rank group after these two sorting operations are ordered by inverse frequency weight of their component words. The weights of the terms are calculated by adding up frequency of occurrence in the thesaurus database of each of their component words (stems). Therefore, if two terms have the same number of matching user terms and made up of the same number of words, the one with a smaller weight is ranked first (see appendix T).

The CILKS system shows the first 30 terms in the ranked list to the user. In the KC, the same limit is applied in testing the thesaurus terms for their 'relatedness' to the user's query on the grounds that, the top portion of the list contains the most useful terms for selection, and it is desirable not to cause unnecessary computing load in cases where large number of thesaurus terms match the user input terms (for queries with high frequency terms in the Inspec thesaurus such as, systems, computers, data, power, and so on, there could be several hundred matching terms in the thesaurus). The matching thesaurus terms for the queries used in experiments 1 and 2 can be found in appendix B and H. Since preferred terms of the lead-in terms are also included, in many cases there are more than 30 terms in the lists. Also note that in appendices B and H the matching top terms are presented in alphabetical order.

This acts as first order filtering of the thesaurus terms matching the query statement (see appendices B, and H). The next level of filtering applied by re-ordering the terms in the list by calculating the *relatedness* of the individual terms to the query terms as a *whole* (appendices C, and I).

7.3.1.3 Ranking of the matching thesaurus terms (Re-order the ranked list according to the value of relatedness):

The objective of this process is to determine the value of relatedness of each term in the list produced in **step C** in section 7.3.1.2 to the whole of the user's query. The formulae used in following systems are identified as suitable candidates in the context of the present project (see 2.2.1) to determine the value of relatedness of the thesaurus terms to the whole of the user's query: AID (Doszkocs, 1978), and LEXIQUEST (Vickery & Vickery, 1993, pp. 127-128).

There are of course numerous methods which calculate the similarity between terms ('term clustering' methods reviewed in section 2.1.2), however they were not considered for the present purpose as we are interested in establishing similarity with thesaural terms and the search statement as a whole and not individual terms in the query, as noted above.

Another obvious candidate for this purpose is the 'Robertson-Sparck Jones' probabilistic model (Robertson & Sparck Jones, 1976; cf. 7.2.2), used in Okapi for calculating the weights of search terms and selecting terms from the relevant documents.

Although it is formulated for the above stated purpose, there is some evidence that it can be used for term selection when there is no relevance data (Robertson et al., 1996, p. 84; see further below).

In the absence of definitive evidence or criteria to base the decision to select one of the above mentioned 'measures of relatedness', it is decided that the Robertson-Sparck Jones formula (will be referred as F4 hereby) is to be used on the grounds that it is already implemented in Okapi, therefore can easily be modified for the purpose of this project to calculate the *relatedness* between the query terms and thesaurus terms.

The derivation of F4 is given in 7.2.2. The following information is required as would be recalled from 7.2.2 to calculate the F4 'selection value' of the term t, w_t:

$w_t = \log\{(r+0.5)/(R-r+0.5)\}/\{(n-r+0.5)/(N-n-R+r+0.5)\}$

where;
n: number of postings for the term t (number of documents indexed by t)
R: number of documents chosen as relevant by the user
r: number of relevant documents containing the term t
N: is the total number of documents in the database, which is a fixed number

While F4 is effective in determining the weights of user input search terms, a modified version of it (Robertson, 1990) is suggested to increase the effectiveness of this method in selecting terms from relevant documents. This is known as 'WPQ selection value'. (ibid).

The ranked list generated as described in 7.3.1.2 is likely to contain many irrelevant terms as, some of the terms in the list would likely to match only one of the user input terms or its stem. This suggests that the problem of ranking terms in *step C* of section 7.3.1.2 is similar to selection of terms from a source which provides indirect evidence (such as documents selected by the user in relevance feedback), rather than from a source that provides direct evidence (such as user input search terms).

As the initial selection of the source terms has a direct and strong effect on the subsequently found linked terms, it is clear that selection of correct source terms is crucial for the success of the rest of the procedure. These considerations have led us to use WPQ instead of F4 in computing the *relatedness values*[1] of the matching thesaurus terms to the user's query (see also further below and section 7.3.2.3).

The modified form of F4, WPQ, is likely to yield better results in predicting relatedness of thesaural terms to the user's query terms (cf. 8.3 and 7.4). It is denoted here as a_t , and given as:

$a_t = w_t\,(p_t - q_t)$

where;
$p=r/R$
$q=(n-r)/(N-R)$
and, r, R, n, N are as defined above.

[1] More accurately, WPQ/F4 calculates the 'probability of relatedness' between the thesaurus terms and the user input search terms.

As mentioned in 7.2.2, WPQ has been derived for selecting terms contained in the documents chosen as relevant by the user. To use this formula (or equally F4) to choose terms from the thesaurus prior to relevance information, it is necessary to modify it. Some researchers have experimented with using top portion of a ranked list of documents to calculate the initial probabilities when there is no relevance data available (e.g. in the first search iteration.) This is sometimes called as the 'top-document feedback' (Fitzpatrick & Dent, 1997). The assumption here is that for a reasonably effective system, precision at the top portion of a ranked list is quite high, therefore, the documents in the top portion of the list could be considered as relevant. There is some evidence from the TREC experiments that, top document feedback does seem to work (Robertson et al., 1996; Buckley et al., 1995)

In WPQ/F4, *N* and *n* are fixed for a given database, however, *R* and therefore *r* are not known. Since we want to establish the relatedness of a thesaurus term to the whole of the user's search terms, it may be assumed, as discussed above, the relevant documents (i.e. *R*) are those indexed by the original user input search terms. This can be called as 'pseudo-relevance'. In this case, *r* can be easily calculated as the subset of documents retrieved by the user's search terms and indexed by the term *t*.

However, as Okapi is a partial match system, queries containing very general terms, such as "computers", "system" and such, would retrieve a list with several thousand documents, most of which do not match the user's query well. The same problem will happen with query statements containing several moderately general terms, such as; "expert systems data base management models". One important feature of Okapi is however to rank the document set in the order of decreasing probability of being relevant to the user (7.2.2). Thus, it should be possible to avoid the above described problem, at least partially, by limiting the retrieved set to those documents at the top portion of the ranked list as suggested by the top-document feedback argument mentioned earlier.

It is worth noting that whereas WPQ takes into account the frequency of the term *t* in *R* in determining its value, F4, *practically*, does not[2]. F4, therefore, assigns higher value to terms who have a very small *n* (or index no documents in the database, i.e. *n*=0) which would bias the selection towards rare terms that do not index any documents in *R*. F4 is therefore very likely to select infrequent thesaurus terms of no relation to the user's search. WPQ corrects this by considering the frequencies in *R*, therefore, it is better suited to select terms from heterogeneous collection of terms, such as the list of thesaurus terms created by the method described in 7.3.1.2 above.

WPQ[3] is therefore chosen to calculate the relatedness value of the thesaurus terms found applying the procedure described in 7.3.1.2 to the user's query statement (see also 7.3.2.3). This is done by taking *R* as the documents at or above the 'weight hump' calculated as some percentage of the total weight of the user's query terms.

Two different systems are designed with slightly differing overall objectives, namely, *KBS-1* and *KBS-2*. The design objectives of both system is discussed in section 7.3.2.3, it is suffice for the purpose of this section to note that *KBS-1* has an edge towards pointing out domains of inquiries

[2] It does, however, the effect may be overwhelmed by the *n/N* component.

[3] As N is a very large number (314427 in Inspec.B collection) compared to R, r and n, and R is constant for a given query, WPQ can be taken as approximately equal to WPQ=(W_t*r), which is used in the actual calculation of the relatedness values in the implemented systems.

that are likely to be *new* to the user (i.e. not foreseen by the user's original search terms), while *KBS-2* is more conservative in its design and biases towards the user's initial search area. This difference in the design of the two systems prompted different estimates for the value of R in equation a_t above.

In *KBS-1*, the weight hump which determines the value of R is taken to be the arithmetic average of the total weight of the user's query terms. In the pilot study (8.2.1.3), it is observed that when the sample size (R) is small it tends to bias towards the terms in the user's query statement with higher WPQ/F4 weights, while represent badly or not at all other terms with relatively low weights that may be present in the query statement. To correct this, a minimum value of 500 is decided for R. When R is below this minimum level at half of the total weight, it is lowered to the next weight hump until the limit is reached or exceeded. Both of these limits are rather arbitrary, suggested merely by the experience gained in the pilot study previous to formal experimentation, and meant to serve as *heuristics* when there is no other reliable guide. Furthermore for two term queries it is decided that, to represent both terms fairly R should be taken as union (logical OR) of the sets indexed by each term in the user's query statement.

The results of the *Experiment 1* indirectly suggest that (cf. 8.3.1.4) an increase in precision can be expected with a lower value for R. Since *KBS-2* aims to have a higher effectiveness, it is decided that R should be lowered to a value corresponding to two-thirds (2/3) of the total weight of the user's query terms. Similarly, the minimum set value for R is lowered from 500 to 300 in *KBS-2*. For two term queries R is taken as union of the sets indexed by each term in the user's query statement when the number of documents in the database indexed by both terms is less than the minimum limit of 300. When the number of documents indexed by both terms is greater or equal to 300, R is taken as the corresponding value.

To illustrate the above described procedure consider one of the queries from experiment 1 (cf. appendix V): "*tracking noise edge*". The weights for these terms as calculated by WPQ (equivalent to F4 at the beginning of the search; cf. 7.2.2) in the Inspec database (inspec.B), are: 56, 53 and 58, respectively. There are 6302 documents in the database indexed by 'track', stem of the term 'tracking' (cf. 7.2.2), 7743 by 'nois', and 5625 by 'edg'; altogether 18687 documents indexed by at least one of the query terms. The arithmetic average of the weights of the query terms is *167/2=83.5*. Therefore, in calculating R, the threshold value of document weight is taken as *83*. Any document below this value is excluded from R. It meant in this particular example that, any document indexed by any two or all of the search terms (i.e. at the weight humps of 167, 114, 111, or 109) is included in R, and any document indexed by only one of the search terms (i.e. at weight humps of 58, 56, or 53) is excluded. This results in R of *976* documents for the query cited above. When the resulting R at 1/2 of the total weight (2/3 in *KBS-2*) is less than 500 (300 in *KBS-2*) the weight hump is lowered to the next level until R is equal to or greater than the minimum level.

The following example from experiment 1 should illustrate this. The user's query statement was: "*texture detection fractals*". The WPQ/F4 weights for these terms are 82, 50 and 89, respectively. The total weight of the user terms is 221. Total number of documents (R) at or above the weight hump of 110 is 251 in the test database. Since this value is less than the minimum required value of 500, the weight hump is lowered to the next level of 89. This results in R of 797. Since this is greater than minimum limit of 500, it is used in the calculation of equation a_t.

7.3.1.4 Selection of the source terms (Select the source terms from the re-ranked list:)

The objective of the matching sub-system of the KC is to provide the terms that will be subsequently used in the term linking sub-system of the KBS (fig. 7.3), which generates the batches of linked-terms noted at the beginning of the present section (7.3.1).

The linked-terms that comprise the batches are generated by using a 'constrained spreading activation' method (7.3.2.1) used in AI applications (Chen, 1992, see 2.2.2), which requires at least an inception (source) node and some constraints to terminate the propagation in the net (Cohen & Kjeldsen, 1987).

Although a single node is enough to apply the method, it is desirable to have more information about the nodes so that it would be possible to constrain the spreading activation without imposing artificial conditions. In many applications, it is therefore expected to have two or more source nodes to benefit from the method (cf. 2.2.2). The source nodes are either user selected, or derived from some information available to the system regarding the user's query or information need.

In the present study it is hypothesised that it is necessary to have at least two source terms to obtain a reasonable description of the user's initial domain of inquiry. The number of source terms more than two however are left out of the present project on the grounds that: *a)* it will be computationally more complicated *b)* it will make the evaluation process more complicated (cf. 8.2), *c)* users of Okapi frequently enter two or three search terms to describe their query (cf. Goker, 1994), therefore, it is unlikely that the search statement contains more than two distinct concepts that can be represented by more than two thesaurus terms (cf. 7.4.1).

An individual batch describes, as noted in the preceding sections, part of the document collection related to the part described by the user's search statement. From this argument it can be further reasoned that, the source terms which are used in the generation of the batches should also selected according to their relatedness to the whole of the user's query as evident from the search statement, rather than relatedness measured between the individual search terms or a group of search terms that describe a concept by forming a phrasal group and the individual thesaurus terms. In any case the second alternative is not feasible in the context of Okapi for the reasons discussed in 7.3.1.1.

The two source terms used in the generation of batches are selected by the procedure described below:

Once the matching thesaurus terms of step C (section 7.3.1.2) are re-ordered following the procedure described in 7.3.1.3, source terms that are used in generating the term clusters (batches of linked terms; cf. 7.3.2, in particular 7.3.2.1 and 7.3.2.3) are selected. The following heuristics are applied in selection of the source terms:

I) Determine the exact matches: This heuristic defines the condition of 'exact match' between the query terms and the thesaurus terms. Three main types of *exact matches* (*I.i* to *I.iii*) are defined below.

The first type as defined by heuristic *I.1* is the only type of exact match which seeks to establish character by character correspondence between the user input search terms and the thesaurus terms in the KB. In this 'strong' sense of the term exact match, each and every user input term is contained by the thesaural term as exactly in the form typed in by the user and there is no other term present in the thesaural term that is not present in the user's search terms.

Heuristics *I.ii* to *I.iii* below provides 'weaker' definitions of exact match by allowing one or more of the following conditions to be present in the definition of the exact match: a) the exact matching thesaural term may not contain all the user input terms b) it may contain terms that are not present in the user's original query c) it may match with the stemmed form of the user's search terms rather than the form as it appears in the user's query.

In exact match type *I.ii*, extra terms not present in the user's query are allowed in the thesaural term if there are no other term in the KB that match with any of the user input terms or their stems.

Even if a thesaural term does not contain all user input terms, it may qualify as an exact match according to the definition of type *I.iii-a* below, if all the user's search terms not represented by it are covered by some other terms in the KB. Exact match of type *I.iii-b* complements type *I.iii-a* by stating that, even if a thesaural term does not contain all user input terms and contain extra terms that are not present in the user's original query, it may qualify as an exact match, if all the user's search terms not represented by it are covered by some other terms in the KB and there is at least one user input search term (or its stem) contained by the thesaural term which is not contained in any other term in the KB.

The following rules of thumb are applied to determine the types of exact match from *I.i* to *I.iii*:

I.i) If a thesaurus term (preferred or lead-in; see 7.1.1) contains *all* user input terms as exactly they are typed-in by the user (not necessarily in the same order as they are input by the user) *and* no other, it is considered as an exact match.

This could be illustrated with a real example from Okapi user logs in which the user input search term consists of only one word, "hypertext", and the matching Inspec thesaurus term, which is a lead-in term, is "hypertext". In this particular case the preferred term for "hypertext" in Inspec thesaurus, "hypermedia", is taken as the exact match of the user's input.

However, in Inspec thesaurus some lead-in terms have more than one preferred term (one has 13, cf. 7.1.1). In such a case all preferred terms of the exact matching lead-in term are taken as exact matches. This rule is applied to all heuristics from *I.i* to *I.iii-a&b.*

Another example of this type exact match occurs for the user input of "database management systems": all the search terms (and no other) are contained by a single compound thesaurus term of "database management systems", which is a preferred term. Note also that in *I.i* type exact match, the input terms themselves and not their stem *must* match with the terms that make up the thesaurus term (and not with their stems), however the order (sequence) of terms in the user input and the compound thesaurus term is not considered as pertinent.

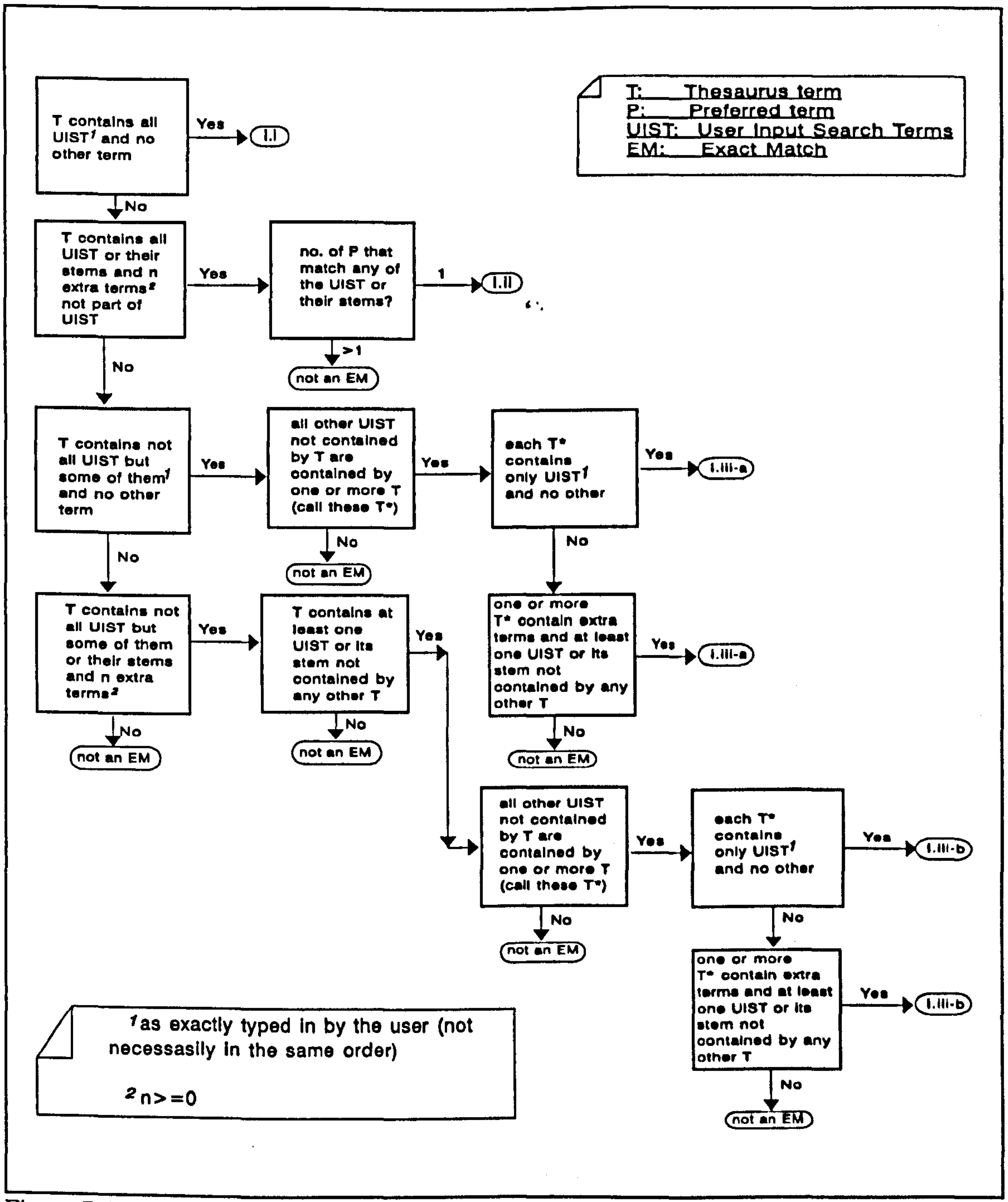

T: Thesaurus term
P: Preferred term
UIST: User Input Search Terms
EM: Exact Match
T contains all UIST¹ and no other term
Yes
I.I
No
T contains all UIST or their stems and n extra terms² not part of UIST
Yes
no. of P that match any of the UIST or their stems?
1
I.II
>1
not an EM
No
T contains not all UIST but some of them¹ and no other term
Yes
all other UIST not contained by T are contained by one or more T (call these T*)
Yes
each T* contains only UIST¹ and no other
Yes
I.iii-a
No
not an EM
No
one or more T* contain extra terms and at least one UIST or its stem not contained by any other T
Yes
I.iii-a
No
not an EM
No
T contains not all UIST but some of them or their stems and n extra terms²
Yes
T contains at least one UIST or its stem not contained by any other T
Yes
No
not an EM
No
not an EM
all other UIST not contained by T are contained by one or more T (call these T*)
Yes
each T* contains only UIST¹ and no other
Yes
I.iii-b
No
not an EM
No
one or more T* contain extra terms and at least one UIST or its stem not contained by any other T
Yes
I.iii-b
No
not an EM
¹as exactly typed in by the user (not necessasily in the same order)
²n>=0

Figure 7.4: Heuristics *I.i* to *I.iii-b*

I.ii) If a thesaurus term (preferred or lead-in) contains *all* user input terms (not necessarily in the same order as they are input by the user) *or* their stems, however may also contain extra term(s) that is/are not in the original user input search terms *and* there is no other *preferred term* in the thesaurus that matches with any of the user input terms (or its stem), it is considered as an exact match.

This type of exact match can be illustrated by the following hypothetical example where the user search statement consists of the word "prominences". There are two terms in the thesaurus that match with this term; "prominences (solar)" which is a lead-in term, and "solar prominences" which is a preferred term. The preferred term for "prominences (solar)" is "solar prominences", therefore, there is only one preferred term in the thesaurus containing the term "prominences" or its stem, thus, an exact match between user input "prominences" and thesaurus term "solar prominences" is established.

I.iii-a) If a thesaurus term (preferred or lead-in) contains not *all* of the user input terms, but some of them as they are exactly typed-in by the user (not necessarily in the same order as they are input by the user) *and* no other term, *and* all the other terms in the user input are contained by some other term(s) in the thesaurus (as they are exactly typed-in) which do/es not contain *any* other extra term *or* if there is/are some extra term(s), at lest one of the user input term (or its stem) contained by the thesaurus term in question is *not* contained by any other preferred term in the thesaurus; the thesaurus terms which fulfil the above conditions are considered as exact matches.

To illustrate this type of exact match consider the following example from experiment 1 in which the user input terms are; "expert systems education". There are two thesaurus terms which contain all the user input terms as they are typed-in and no other, namely; "expert systems", and "education". In this particular example therefore, there are two terms in the thesaurus (knowledge-base; 7.1.2, 7.1.3) that make an exact match with the user produced search terms.

I.iii-b) If a thesaurus term (preferred or lead-in) contains not *all* of the user input terms, but one or more of them or their stems, however may also contain some extra term(s) that is (are) not one of the original user input terms *and* there is at least one user input term (or its stem) contained by the thesaurus term in question that is *not* contained by *any* other *preferred term*, it is considered as an exact match, on condition that, all the other terms in the original user input, are covered by some other thesaurus term(s), such that, the thesaurus term(s) containing them is either consists of *solely* the user input terms *and* no other, *or* if there is/are some extra term(s), at least one of the user input term (or its stem) contained by the thesaurus term in question must not contained by any other preferred term in the thesaurus.

As a matter of fact, *I.iii-b* states exactly the same conditions as *I.iii-a* above. However, the conditions formulated in *I.iii-a* is rephrased here for convenience and clarity.

An example from experiment 1 (appendix B) should make this clearer: the user search statement was "polarisation mqw splitter". First of the user input terms, namely, "polarisation" is a preferred term in the thesaurus, there is only one term in the thesaurus which contain the term "mqw" which is "mqw lasers", and it is a lead-in term for which "semiconductor lasers" is the preferred term, and "splitter" matches with the following term, "beam splitters optical", which is a lead-in term that leads to the preferred term "optical elements". In this example, as the query terms "mqw" and "splitter" are contained by two different thesaurus terms (each containing one of the user terms) which do contain some additional terms, however, there is no other preferred term in the thesaurus that matches with either of these two user input terms, and since the query term "polarisation" exactly matches with the thesaurus term "polarisation" which does not contain any extra term, the following thesaurus terms are taken as exact matches: "polarisation", "semiconductor lasers", and "optical elements".

The above four heuristics define the condition of exact match. Selection of the source terms is determined by the number of exact matches as described by the following rules.

II) Select the two source terms: The following rules are followed in the selection of the two source terms used in the batches of terms generation module of the KC:

II.i) If the number of exact matches[4] is one[5], then select the exact matching term. Select a second source term from the re-ordered list (the procedure of which is described in 7.3.1.3) with the highest relatedness value.

II.ii) If the number of exact matches is two, *and* there is at least one differing user input term between the two exact matching thesaurus terms, then select both of them.

II.iii) If the number of exact matches is two or more, *and* there is *no* differing user input term between any of the exact matching thesaurus terms, then select the two exact matching terms with the 'highest relatedness value' (HRV). In the case of a tie, select any one of the terms in tie.

II.iv) If the number of exact matches is more than two, *and* there is only one pair of exact matching thesaurus terms differing in at least one user input term, then select those two exact matching thesaurus terms.

II.v) If the number of exact matches is more than two, *and* there are more than one pair of exact matching thesaurus terms differing in at least one user input term, then select the exact matching thesaurus term with HRV as one of the source terms. Select the second source term among the exact matching terms that *complements* the first source term (i.e. has got at least one user input term that is not contained by the first source term) with the highest relatedness value (HRV). In the case of a tie, select any of the terms in tie.

II.vi) If there is *no* thesaurus term in the knowledge-base that makes an exact match with the user input, then select the thesaurus terms with the highest relatedness value as the first source term. Select the second source term among the terms that *complement* the first source term (i.e. has got at least one user input term that is not contained by the first source term) with the highest relatedness value. If there is no thesaurus term that complements the first source term, select the one with the highest relatedness value. If there is a tie select any one of the terms in tie. The weights and ranks of the terms matching the queries used in experiments 1 and 2 are given in appendices C and I respectively. Note that only preferred terms are ranked. As lead-in terms do not index any documents, *r* in F4/WPQ becomes zero automatically.

[4] *Exact match* as defined above through heuristics *I.i* to *I.iii-a&b*.

[5] Could only be type either *I.i* or *I.ii*.

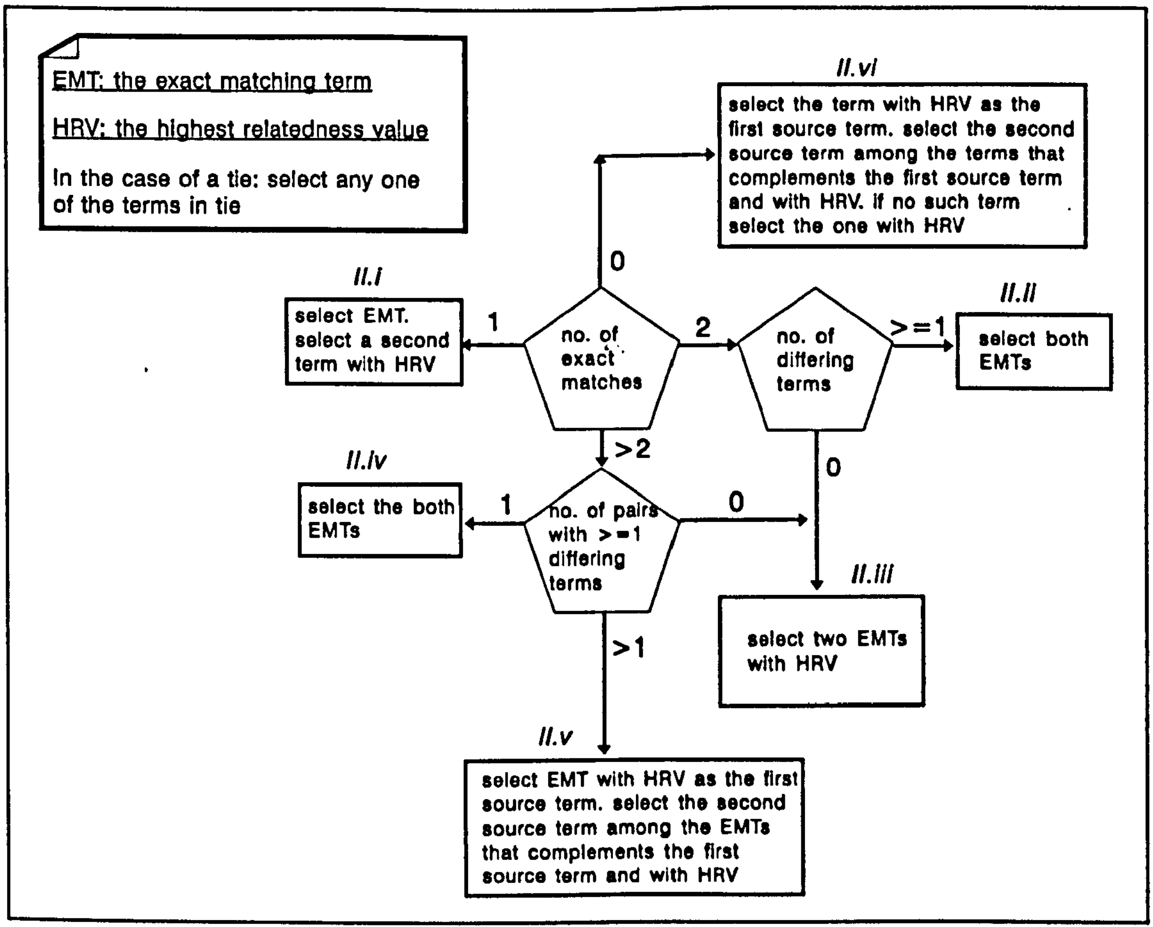

EMT: the exact matching term
HRV: the highest relatedness value
In the case of a tie: select any one of the terms in tie
II.vi
select the term with HRV as the first source term. select the second source term among the terms that complements the first source term and with HRV. If no such term select the one with HRV
0
II.i
select EMT. select a second term with HRV
1
no. of exact matches
2
no. of differing terms
>=1
II.ii
select both EMTs
>2
II.iv
select the both EMTs
1
no. of pairs with >=1 differing terms
0
0
II.iii
select two EMTs with HRV
>1
II.v
select EMT with HRV as the first source term. select the second source term among the EMTs that complements the first source term and with HRV

Figure 7.5: Heuristics *II.i* to *II.vi*

III) Test the selected source terms for the distance between them: An additional constrained is applied in the selection of the source terms for the spreading activation used in generation of the batches (linked thesaurus terms) described in the next section (7.3.2). It is hypothesised that if the number of nodes in the knowledge-base (cf. 7.1.3) separating the two source terms is less than one, they are too close to each other to describe adequately the knowledge space (subject/s) implied by the user's query terms. Similarly, it is assumed that if the number of nodes between the two source term is more than 6, they are too distant from each other to describe adequately the user's search space. This latter assumption is generally made with other systems that make use of the semantic network and spreading activation methods (see 2.2.2). Therefore, the following rule of thumb is adopted:

III.i) If the number of terms (nodes) between the two source terms is less than 1 or more than 6, keep the thesaurus term with the least number of extra terms (terms not contained in the user's query). If both thesaurus terms have the same number of extra terms, keep the one with more user input terms. In the case of a tie in the number of matching user input terms select the one with a higher relatedness value. If this does not resolve the tie, select any of them. Discard the other thesaurus term. Go to step *II* and select another thesaurus term to replace the discarded source term. Repeat step *II.i* to *II.vi* until the second source term is found.

7.3.2 Linked terms

Once the source terms are selected by following the procedure given in 7.3.1, they are used for the purpose of generating batches of terms linked in the thesaurus. The hypothesis in generating batches or clusters of terms linked by one of the relationships in the thesaurus (cf. 7.1.3) is that, terms that are connected in the thesaurus by some criteria represent the same underlying concept (Paice, 1991; Chen, 1992; cf. 2.2.2). However, this may be articulated better, in terms of concepts and tools presented and developed in chapters 3 to 6 of this dissertation.

As it would be recalled from 3.5.5 that, value of.a term in a linguistic system or generally in a sign system is determined by its relative position with respect to other terms or signs that constitute a code (cf. 3.5.5, 3.5.9). In short, value of a unit in a signifying system is a product of other units that circumscribe and delimit it. Thus, it is clear that terms in a thesaurus similarly assume value by means of attracting certain terms and repelling others. As in any signifying system a thesaurus term's meaning is determined by its relation to the other terms in the system. It is observed that a thesaural term in isolation from other terms that relate to it, is difficult to judge in terms of its usefulness for a particular search (see 8.4.2 for the discussion of this matter arising in the CILKS experiments).

It is therefore reasoned that, a certain number of thesaurus terms that are related to each other in some meaningful way (i.e. conceptually) is needed to delimit the value of the individual terms in the batches. In other words, it can be inferred from the semiotic point of view presented in the preceding chapters that, it is necessary to put a given term in *context*, by showing it in the company a number of paradigmatically (cf. 3.5.7) related terms that satisfactorily delimit it. This is necessary for both the task of presenting to the user a group of terms that make a conceptually coherent set (i.e. from a cognitive point of view), and to retrieve documents that are represented by a particular sense of the term in question, and not by any other (i.e. from a pragmatic/retrieval point of view). It is assumed regarding to the second point that, it is possible to combine the terms during the search such that they are used together to represent the attributes of the required documents (see 7.3.3 for further discussion of this point).

It can be recalled from section 5.7.2 that, in the encyclopedic model of the universe of knowledge, certain paths of connections in the network of semes (or interpretants in the semantic universe) interpret particular senses of a given sememe. The interpretation of the encyclopedic model in the context of the documentary information retrieval situation of section 5.7.3 makes possible to think sememe as an index term or a group of index terms, and their interpretants or semes are being other index terms. It is thus possible to posit that, user input search terms constitute a sememe and the thesaurus terms that are associated with the query corresponds to its semes; i.e. a set of semes associated with a given sememe at a particular moment, delimiting its particular sense or reading for that instant (cf. 5.7.2; 5.7.1).

The objective of the KBS, as stated in 6.3.2 and 6.4.3, is to find documents related to the those described by the user's search terms without necessarily being like them. Therefore, given that search terms delimit a portion of the document collection, some other portions of the document collection that are related to the portion delimited by the search terms can be circumscribed by batches of terms (in this project from a thesaurus) that are related to the user input search terms.

Since the number of user input terms in Okapi is usually 2 to 3 (Goker, 1994), it can be expected that knowledge space (or part of the document collection) defined by the search terms is often quite large (general). It can therefore be reasoned that knowledge space delineated by adding extra terms (cf. 7.3.3) that are related to -- i.e. co-occur in some documents with -- the search terms is a smaller section of the original domain defined by the user's search terms or a subset of the part of the document collection retrieved by the user's search terms. However, as Okapi is a best match system (cf. 7.2), often only a few of the search terms (in a batch) are likely to be contained by the documents in the retrieved set (cf. 7.3.3 and 7.4.1). Therefore, it is not necessary that an expanded query statement with added thesaurus terms would always be more specific than a query with a few free text terms. For this reason, it is perhaps more accurate to claim that (cf. 6.3.3.2), batches of semantically linked thesaurus terms delineate the parts of the document collection that border the part defined by the user's search terms (cf. fig. 6.3 in section 6.3.1).

Having articulated a theoretical basis for the linked terms (batches) in terms of the analytical framework presented in chapters 4, 5, and 6 of this dissertation, it can now be proceeded with the implementation issues, which is done in the following section.

7.3.2.1 Generation of linked terms

Once it is explicated that a number of thesaurus terms are required both to represent the user's query unambiguously and search the database effectively, the main parameter need to be determined is the number of such terms needed to delimit one sense of the query from another. As users' queries are represented in the KBS by batches of terms related in the thesaurus (cf. 7.3.2), the above question can be re-phrased as "how many terms are needed in a batch to distinguish it from another one?"[6].

The above question can be more conveniently posed as the number of distinct batches required to define/delimitate the different senses of the query. The question then becomes a matter of determining the number of batches needed that circumscribe satisfactorily the part of the document collection that is defined by the user's search terms. The problem is thus one of deciding the number of batches -- that are related yet distinct from each other -- required to circumscribe a portion of the document collection. The relatedness between batches can be reasonably determined by the presence of some common terms, while distinctiveness can be measured by the presence of at least one differing term in each pair of batch.

Having formulated the problem in above terms, some heuristics are required to determine reasonable number of such batches needed to circumscribe the part of the document collection related to the part defined by the user input terms.

The initial *ad hoc* investigations suggested that (cf. 8.2.1.3), when the shortest[7] possible connection in the knowledge-base (7.1) between two source terms is traversed, there is often one

[6] One can equally state the same problem as the number of thesaurus terms needed (in a batch) to distinguish a particular sense of a member of the batch from its other possible senses (or simply, to put the (thesaurus) term in context, cf. 7.3.2). By applying the reasoning of 5.7.2, each member of the batch in its turn becomes a sememe interpreted with all the other terms (semes) in the batch, and then a seme which together with other semes in the batch interprets another term (which becomes a sememe) of the same batch.

[7] That is, the least number of intervening nodes between the source terms.

or a few distinct (i.e. differing in at least one node traversed) such paths[8]. It is indeed relatively unlike that more than 3 to 5[9] separate such paths are generated when the least number of nodes connecting the source nodes are traversed. From the same ad hoc analysis, it looked reasonable to suggest that, when there is a small number (5 to 10) of distinct paths (or batches[10]) connecting given two nodes, they seemed to represent larger chunks of the semantic space that remains ambiguous in cognitive terms. In other words, it seems that small number of batches do not delineate the search space satisfactorily. It is furthermore observed that number of terms in batches that cover the shortest distance between the source terms tend to be small, usually around 3 to 4, including the source terms. It is hypothesised that this is the cause of the difficulty in perception of the batch as representing a singular topic. In short the smaller the number of terms in a batch, the more ambiguous and/or conceptually general they are perceived as a whole.

The two phenomena are of course causally related; larger the number of terms in a batch, i.e. larger the levels of expansion (cf. 7.1.3), larger becomes the possibility of generating distinct paths. The number of distinct batches generated with a pair of source terms is a function of several parameters, such as, the number of terms connected to the source terms, the number of connections each term connected to the source terms have, and the number of connections of the terms at the next level, and so on, as well as, the minimum number of expansion levels needed to connect the source terms. Since, these characteristics of the terms in the Inspec thesaurus vary a great deal (7.1.2) it is difficult to discern a regular pattern that relates the number of batches to the any of the properties of the source terms mentioned above. Therefore, it is not possible to relate by simple formulation the number of terms in a batch to the number of distinct batches. However, as a general rule it can be stated that, higher the number of terms in a batch, higher the number of resulting distinct batches.

It is observed further that, when a single level of redundancy[11] is introduced to the expansion process (spreading activation), the number of unique batches (paths) leaps from around 5-10 to several times of this number. As a simple guiding rule, it this thus possible to state that, when one extra term is introduced to the batch of terms that represents the shortest distance between two terms, i.e. when an extra level is expanded (one redundant term, first degree of redundancy), the number of batches increases substantially (cf. 8.2.1.3, the pilot), which then results in better delineation of the search space, as well as, better perception of the linked terms in the batches in cognitive terms (i.e. semantically less ambiguous, better defined).

However, the above rule of thumb is not always observed, that is, introduction of one redundant term does not always result in substantial numbers of new paths. There may be only a few new paths emerging with one or even two redundant terms (cf. 7.3.2.2). To simplify the matter, it is decided that the heuristic to be followed in determining the number of expansion levels should relate to the number of batches generated. The assumption behind this is that, a part of the

[8] A path being a collocation of nodes and the associated relation types in the knowledge-base (cf. 7.1.3) that link two source terms (nodes).

[9] In the initial investigation (cf. the pilot study in 8.2.1.3) the maximum number of batches encountered at the minimum number of expansion levels needed was 10.

[10] A batch contains the nodes of an individual path.

[11] One more level of expansion than the minimum level of expansion required for connecting the two source terms.

knowledge domain is only well circumscribed when there is quite a few number of batches that delineate it. This is justified by the subjective investigations done in the pilot study (8.2.1.3). It is previously noted that, in the pilot study 10 was the maximum number of batches required to connect the two source terms in the shortest possible way, therefore it seems to be a good candidate for delimiting the boundary between number of batches that satisfactorily map a knowledge space and that defines an ambiguous portion of knowledge space. For this reason, in the design of both systems, it is taken as the boundary condition[12]. The following heuristic thus results:

Heuristics IV: If the number of batches is less or equal to 10 at *l* levels of expansion, expand for a *l*+1 levels.

7.3.2.2 The spreading activation

The role of thesaurus as a semantic network has been reviewed in 2.2.2. Construction of the knowledge-base (KB) used in this project from the Inspec thesaurus has been discussed in 7.1.2. The KB effectively corresponds to a semantic network (cf. 7.1.3).

The KB constructed as a semantic network can be searched by a technique known as spreading activation (Shoval, 1986; 1985; Cohen & Kjeldsen, 1987; Chen, 1992; 1991; Paice, 1991; Croft & Thompson, 1987). Spreading activation is a process of connecting the nodes in a semantic net, such that: first, terms that are connected to the input terms (referred as source terms in this dissertation) are activated -- that is, selected to replaces the input nodes -- which are then used to activate the next level of terms connected with them, and so on, until some condition is satisfied and the propagation in the net terminates. The resulting pattern of connected terms are then used in the selection of the terms for augmenting or replacing the user input terms. Various methods of using this process to select thesaurus terms for IR purposes are described in sections 2.2.2. The activation of nodes is normally subject to some conditions or constraints, in which case the process is usually called as *constrained spreading activation* (Cohen & Kjeldsen, 1987). The type of relations connecting terms in a knowledge-base take various forms (cf. 2.2.2).

In the KC, the spreading activation starts by activating all terms connected with any one of the two source terms[13] (see 7.3.1.4). The types of relations exist in the KB used in this project connecting the terms is discussed in 7.1.3. The terms related to one of the source terms with any one of the available relations in the KB are selected initially. The activation process then progress by activating all terms connected with the terms activated at the previous level of propagation, except those violating *Rule I* (see below). The term which generates the terms at the next level of the activation process are referred to as the 'parent term', and terms originating from it as its 'progeny'. The spreading activation progresses in the KB until the number of levels defined by *Heuristics IV* is reached.

All types of available relations in the KB are permitted in linking a term with another. Only those leading a term back to its parent (or grand-parent!) are excluded:

[12] However, this is not by any means a definite number, it must be stated that it is a rather an arbitrary rule of thumb.

[13] The actual implementation of the search algorithm in Oracle SQL works by initiating the spreading activation from the both ends, i.e. both source terms. See below for more detail on this.

Rule I: Reverse activation in the form of; "$term_A$ --> (activates) $term_B$ followed by $term_B$ --> $term_A$", or "$term_A$ --> $term_B$ --> ... --> $term_A$ --> ...", is not allowed.

Since, all relationships in the Inspec database and hence the resulting KB, have an inverse (cf. 7.1.2), this rule prevents circular propagation in the KB which does not produce new information (term).

The source terms are separated by at least 1 node and at most by 6 in the KB, according to *Heuristics III.i* (7.3.1.4). The minimum number of activation level can therefore be 2, and the maximum level can be 7. When the number of activation levels prescribed by *Heuristics IV* (7.3.2.1) is reached the activation process terminates. The unique paths as defined in 7.3.2.1[14] are then selected to produce the individual batches that represent a part of the knowledge domain (or the document collection), by picking-up the nodes that make-up the paths. Each unique path gives rise to a unique batch with l+1 nodes (including the two source terms), where l is the level of activation reached when the propagation in the KB is finally halted.

One interesting aspect of the spreading activation technique implemented in this project, which is unlike to those implemented in some other systems (e.g. Chen, 1992; see also 7.4.2) is that, it does not discriminate between different types of relations. This is to do with the design objectives of the system. As it would be recalled from 6.3., the design approach to IRS as developed in this project aims to avoid making strong assumptions regarding usefulness of certain documents or terms to the user as much as reasonably can in a given design context. In the overall context of the designed retrieval systems, batches carry far more importance both in the search process and in their function of representing a knowledge domain than the individual terms that make up the batches. Therefore, it makes little sense in the implemented KBSs to discriminate between different relationships that exist in the knowledge-base.

It is worthwhile to note in this connection that, the synonymy (equivalence) relation is of particular interest. The peculiar characteristic of having more than one preferred term of some lead-in terms in the Inspec thesaurus is noted in 7.1.1. This puts the usefulness of the synonymy relation in doubt altogether. Furthermore, a lead-in term picked-up in a path by following the UF (used for) relationship (cf. 7.1.1) is of no use in the retrieval process itself (cf. 7.3.3), which makes doubtful of any merit of tracing the synonymy (UF) relation. However, as long as the possibility of a lead-in term facilitating inclusion of another useful term in the path being traced remains, it is worthwhile not to preclude it. As usefulness of a batch as a whole is judged eventually (cf. 7.3.2.3), any a priori attachment of value to particular types of relation in the KB is disregarded (see 7.4.2 for more on this).

Another point worth noting here briefly is the effect of activating redundant and a priori determined number of levels (rather than following 'the shortest possible distance creed') on the batches generated. The particularly interesting question to be addresses in this regard is: whether or not some of the batches that would have been generated at l activation levels eliminated altogether by processing at l+1 activation levels. In other words, the pertaining question is that: should at l level of activation exist a batch B_l such that, it consists of specific nodes of N_{b1} N_{b2} ... N_{bl+1}, would there be a batch at a $l+1$ levels of activation containing all the elements (nodes) of the batch B_l (not necessarily in the same order of B_l) and one other term? This question and its possible implications for the KBS is addressed in 7.4.2.

[14] Each path therefore starts with one of the source terms and terminates with the other source term.

The last point need mentioning in this section is to do with the actual implementation issues of the spreading activation process in the KC (cf. footnote 13). The details of the spreading activation program used in the project is given in appendix S. It is sufficient to note here that, it starts by activating all terms up to *l/2* levels from any one of the two source terms, where *l* is the total number of levels to be activated which is prescribed by *Heuristics IV* of section 7.3.2.1. At the second half of the process, *l/2* levels are activated from the end of the other source term. The paths containing the same term, each originating from one of the source terms, at *l/2* level of expansion are then joined to form one complete path leading from one of the source terms to the other. If *l* is an even number, activation is applied for same number of levels (i.e. *l/2*) from both ends, when *l* is an odd number, one of the source nodes are expanded for *m* and the other for *n* levels, such that $m+n=l$ and $m-n=1$.

7.3.2.3 Ranking of batches

The batches are generated such that, certain amount of redundancy in the number of terms to represent the original user's query is built in. In other words, the expansion level in the generation of batches is such that, it is usually (but not necessarily)[15] more than the minimum number necessary to connect the two source terms. The reason for the introduction of redundancy is to find the documents related to but not necessarily prescribed by the user's search terms, as discussed in 6.3.2.

However, once the batches of terms describing such documents are produced, it is desirable to rank them in terms of their relative value of relatedness to the user input search terms for convenience as noted again in 6.3.2. This is a desirable thing in terms of the adaptability of the retrieval process to different user needs, especially when considered that Okapi's user population varies a great deal in their expectation of performing the prescriptive or denotative, inventive or reproductive types of retrieval games (cf. 6.3.2).

This section deals with the process of ranking of the batches whose generation is discussed in 7.3.2.1 and 7.3.2.2.

The objective of ranking of batches, as stated above, is to calculate the relatedness value of each batch as a whole to the whole of the user's search terms. The discussion of how to measure the relatedness of individual thesaurus terms to the user's search terms as a whole has been presented in 7.3.1.3. The present discussion is in essence similar to the to that of 7.3.1.3.

As noted in 7.3.1.3, two different systems are designed with slightly different overall design objectives. Two experiments are set-up to evaluate the two systems (8.2). The first system,

[15] *Heuristics IV* prescribes that, the number of unique paths between the two terms must be more than 10. Although this does not in itself prescribe actualization of a redundant level(s) of expansion, in practice all cases in the evaluation tests (cf. 8.3) had at least one (at times up to 3) redundant levels of expansion to meet the condition laid down by *Heuristics IV.* Redundancy is of course measured in comparison to the *least number* of expansion level required to connect the two source terms. If the number of level of expansions to connect the source terms is two, one level of redundancy is achieved by expanding for three levels from one of the source terms or, as in the actual implementation of the spreading activation process in *KBS-1* and *KBS-2,* two levels from one of the source terms and one level from the other, joining at the common terms found by the two paths traversed from both ends.

KBS-1 has the objective of playing the inventive (prescriptive) retrieval game. That is, *KBS-1* aims to point out to the user those documents and knowledge spaces that are not immediately obvious from the original user input search terms. The corresponding experimental set-up therefore aims to elicit information regarding this objective (8.2.1.1). The second system, *KBS-2*, has a slightly different aim than *KBS-1*. It still does aim to play the prescriptive game, however, its design objective is, relative to *KBS-1*, slightly more conservative, in that, it is concerned *more* with finding the documents as described by the user's original query terms. It therefore aims to increase precision while also suggesting documents that are not foreseen by the user input terms. To do this, it is designed to attach more importance to the thesaurus terms that are more *likely to be related* to the user input terms. *KBS-1* on the other hand is designed to bias towards terms that are *less likely to be related* to the user's search terms. One way of achieving this as applied in *KBS-1* is to use a weighting formula which is less biased towards terms contained by the documents chosen by the user (i.e. those indexed by the user's original search terms, which correspond in KBSs to (pseudo-)relevant documents).

In our interpretation of F4/WPQ formulae, relevant documents chosen by the user are taken to be a certain portion of the documents indexed by the user input search terms (7.3.1.3). WPQ takes into account the frequency of a term in the relevant documents more strongly in calculating its value (cf. 7.2.2), while this is reflected weakly in F4 weights. It is decided hence that in *KBS-1* F4 should be used to rank the batches rather than WPQ in order to increase the bias towards batches that contain thesaurus terms that are not directly related to the user input terms, so that, it would fulfil its design objective of suggesting to the user new terms, new areas of inquiry. As *KBS-2* aims to increase precision (in comparison to Okapi, cf. 8.2.1.2), in *KBS-2*, WPQ is used after the results of experiment 1 (see 8.3.1.4, and appendix P) suggested that WPQ was better in ranking the batches in terms of precision values (see below).

F4 weights have the property of being simply added up to calculate the total weight of a document indexed by them (Robertson & Sparck Jones, see also 2.1.1 and 7.2.2). In other words, simple sum of F4 weights of query terms is used to calculate the total weight of a document indexed by those terms. It is therefore relatively straightforward to suggest that the overall value of relatedness of a batch to a query can be similarly calculated by adding up the F4 weights of the individual terms that make up a batch. WPQ values, unlike F4 weights, do not have the property of being simply added up to calculate the total weight of a document indexed by them. It is decided to use the relevance judgements of the users in experiment 1 to test the ability of WPQ in calculating the batch weights by adding up the WPQ values of the individual terms comprising the batches. As the same information can also be used to test other formulae, it is decided to test also the ability of formulae used in AID and LEXIQUEST (identified in 2.2.1, cf. also 7.3.1.3) in ranking the batches (8.2.2.1 and 8.3.1.4, see also appendix P).

It is worth remembering that, in both systems, selection of the source terms are done by WPQ, as it is essential for both systems to find the source terms that are likely to be of relation to the user input search terms, so that, subsequently generated batches of linked terms which depend heavily on the accuracy of the initial selection of the source terms have a good chance of not being diverted to areas which are totally unconnected with the user's original query. It is worth realizing that, selection of the source terms are made from a list which contains terms that only appear there by virtue of containing as little as one user search term or its stem (cf. 7.3.1.2). It is extremely likely that, many of the terms in the initial list from which the source terms are selected are totally irrelevant to the user. Therefore, it is of vital importance to choose the right terms from the list on which the rest of the process of generation of the batches of linked terms depend totally. For this job WPQ is better suited than F4 (cf. 7.2.2) and used in both KBSs.

The parameters in F4/WPQ and how they are interpreted in this project in the context of estimating the relatedness values of the thesaurus terms to the user input search terms have been discussed in 7.3.1.3, including the minimum limit of 500/300 documents for the sample size of *R*. The very same method described in 7.3.1.3 in estimating the parameters of F4/WPQ is used in calculating the weights of the individual terms that make up a batch. In sum, in *KBS-1*, F4 is used to calculate the weight of each individual term in a batch with a minimum value of 500 for *R*. In *KBS-2*, WPQ is used in the calculation of the weights of the terms in batches with a minimum value of 300 for *R*.

The *total* weight of a batch ('batch weight') is then found by adding up the weights of the individual terms that make up the batch. The batches are then ranked according to decreasing probability of being related to the user's query by listing them in descending order of batch weights of F4 and WPQ for the systems *KBS-1* and *KBS-2*, respectively (see appendices D and J for the weights of the top 10 batches used in experiments 1 and 2, respectively. See appendix U for the various scripts used in weighting the thesaurus terms and batches).

7.3.3 Query re-formulation and searching

Once the thesaurus terms that are likely to be of use for the purposes of the retrieval systems *KBS-1* and *KBS-2* are determined and grouped in batches such that each batch starts and ends with the same two term (the source terms) that are highly likely to be related to the user's original search terms, while differing at least in one term, they are used in re-formulation of the user's query statement.

The purpose of the query re-formulation in this project is, in general terms, to let the user of the retrieval system to perform the prescriptive retrieval game (cf. 6.2.2.2; 6.3) in which the user would like to see documents related to her general domain of enquiry, however, further specification of the sort of documents required is either not known (not prescribed yet as in the case of moderate or radical inventions; cf. 6.2.2.2) or unavailable due to the user's difficulty in formulating the search statement. Possible reasons for this difficulty have been discussed in 6.2.2.2 and 6.3.1, and include common problems such as, lack of knowledge of the domain of inquiry and its vocabulary.

It is further hypothesised that, user's search terms define a portion of the knowledge domain constituted by the document collection, and this domain (being often quite general in Okapi searches, since users' input usually consists of 2 to 3 terms on average, cf. Goker, 1984) can be delineated by projecting the user's search terms onto the thesaurus terms (thesaurus is viewed as a map of the knowledge domain, Paice, 1991, p. 436; cf. 7.3.2) such that, collocation of related thesaurus terms (a batch in the terminology used here) represents, as a whole, part of the knowledge domain related to the part defined by user input search terms (cf. 7.3.2.1). It can be assumed that the area defined by a batch is usually more specific than the one defined by the user's search terms, as each batch in this project contains several (on average 5 to 6) thesaural terms compared to 2 to 3 free text terms in an average user input[16] (cf. 7.3.2.1).

[16] However, due to best match searching this effect virtually diminishes as one goes down the ranked list.

A number of such batches are needed to delimit the domain related to the user's query (7.3.2.1), so that the user can investigate (search) different parts or aspects of the domain.

The degree of circumscribing the domain, i.e. to what extent different parts of the domain are made visible, in other words, the extent (detail) of the portion of the map of the domain made available to the user, depends on the terms included (and therefore, excluded) in the batches, which can be controlled by adjusting the parameters of the system. For example in *KBS-1*, it is expected that batches containing unlikely terms in relation to the user input search terms are represented better compared to *KBS-2*. Thus, *KBS-1* shows parts of the knowledge domain relatively far afield from the part described by the user input terms by using a slightly different weighting algorithm.

The eventual search statement to retrieve the documents using Okapi's search engine (cf. 7.2.2) is constructed by combining the user input terms with all the terms in a batch. The reason for combining the terms in a batch with the user input terms is to attain a better retrieval effectiveness, as well, as achieving a better representation of the user's query (see section 7.4.1 for this point). The effectiveness of combining thesaurus terms with user input terms is suggested both by the pilot study in this project (cf. 8.2.1.3), and the CILKS study (Jones et al., 1995).

Although *KBS-1* and *KBS-2* have different design objectives in terms of retrieval effectiveness (cf. 7.3.2.3), it is decided that in both systems it is desirable to include user search terms to attain reasonable level of performance. This is to do with the design objectives common to both systems. Both systems are designed to achieve a balance between conflicting goals of novelty and effectiveness, i.e. suggesting to the users those documents that are not predicted (well or at all) by the original user search terms and predicting those that have higher probability of resembling to the ones described by the original user input terms (cf. 6.3.1 and 6.3.2). Although the degree of this balance varies in the two systems slightly as discussed previously.

The query is formulated using Okapi's search system in the following manner: user's search terms (free-text) are taken as they are typed-in and processed using Okapi's standard procedure (cf. 7.2.2); all of the thesaurus terms in a given batch are taken (except those mentioned below) and combined with the free text terms. In the search process, the thesaurus terms are restricted to the 'descriptor terms field' of the Inspec records (cf. 8.2.2.4), while free text terms are searched in all searchable fields of the records except the descriptor field. In searching, weights for both free text and thesaurus terms are calculated using a variant of F4 which is known to be performing particularly well (i.e. bm11; Robertson & Walker, 1994).

In *KBS-2*, thesaurus terms with the status top (cf. 7.1.2) are excluded from the search statement to increase the *precision* (this is one of the design objectives of the *KBS-2*; cf. 7.3.2.3) after the findings of experiment 1 (cf. 8.3.1) which suggests that top terms are often undesirable to the users and they tend to reduce precision.

The lead-in terms are omitted from the search statement in both systems as they do not index any documents. It would be recalled from 7.3.2.2 that, the lead-in terms are there only as a means of reaching other terms in the net.

7.4 Discussion of the design parameters

The following sections are intended as a further clarification of some of the design parameters and decisions by discussing them in relation to the design objectives and goals articulated in chapter 6.

The main topics discussed are: representation of user's query by thesaurus terms (7.4.1), processes of generation and ranking of the batches (7.4.2), and searching of the document collection (7.4.3).

7.4.1 Representation of the user's query

In the discussion of selection of source terms (7.3.1.4) it has been noted that the design of the *KBS* (both *KBS-1* and *KBS-2*) is restricted with regard to the number of source terms used in the generation of the batches (representation of the user's query).

The main function of the source terms is to map the user input terms to some controlled vocabulary (thesaurus terms). A thesaurus term is effectively a category or a concept in the Aristotelian sense of the term (Andersen, 1986). Therefore, the whole procedure of mapping the user's search terms onto thesaural terms, is one of a classification.

The batches derived from the source terms can be viewed as representations of parts or facets of the subject domain defined by the two source terms[17]. It is therefore important as a design matter to decide on what concepts are present in the user's query and how to represent them. More accurately, it is a matter of design decision to determine what number of concepts should be used to best cover (represent) a given query or how to classify it.

In the particular design approach developed in this dissertation the number of concepts to represent a given query statement is limited to two for the reasons explained in 7.3.1.4. The implications of this with regard to representation of the users' query is that, if there are more than two distinct concepts in the user's query, some of them may not be represented[18] in the batches at all, therefore in the final query statement.

As it could be recalled from 7.3.1.4, four types of exact match conditions are defined. Only in types *I.i* and *I.ii* all the terms input by the user are contained by a single thesaurus term. In all the others there may be more than two exact matching thesaurus terms of which only two are to be selected. In this case some of the user's search terms may not be covered by either of the two selected source terms. The query cited in 7.3.1.4 from experiment 1 ("polarisation mqw splitter") form an exact match of type ***I.iii-b*** such that, there are three exact matching thesaurus terms, namely; "polarisation", "semiconductor lasers", and "optical elements". In this particular case, the thesaurus terms "polarisation" and "optical elements" were selected as source terms. The third exact matching thesaurus term "semiconductor laser", which represents the user input

[17] This was the main interpretation of the batches in sections 7.3.2 and 7.3.2.1.

[18] Representation in this context means that, the user input term itself or its stem is contained by some preferred term in one of the batches. If the user input term (or its stem) is contained by a lead-in term and the lead-in term itself or any of its preferred terms is present in one of the batches, it is also considered that the user input term is represented in the final search statement.

term "mqw" was not chosen as a source term and in none of the batches generated from the two source terms selected, neither it, nor an equivalent term for it appears. Hence, one of the user's search terms, "mqw", was not represented in the final query statement at all. The results for this search was particularly poor (cf. 8.3.1), reinforcing the suspicion that omission of a user's search term from the batches altogether may cause poor performance.

This problem could also occur when there is no exact matching terms as defined by heuristics *I.i* to *I.iii-a&b*. When there is no exact matching terms, the source terms are selected according to the heuristic *II.vi* of section 7.3.1.4. This heuristic states that, the two matching terms from the initial list (cf. 7.3.1.2) with the highest relatedness values (cf. 7.3.1.3) that complement each other are to be selected (cf. 7.3.1.4). In such a case, it could be very well that some of the user input terms are not represented in the source terms at all, and consequently in the batches generated from them[19]. Actually, only types *I.i* and *I.ii* guarantee that all of the user's search terms are represented in the batches. In all other possible cases, some of the user's terms risk not to be represented altogether.

It is important to realize therefore that, number of the source terms used in the generation of batches has a direct affect on the representation of different concepts or aspects that a user's query may contain.

In the practicality of searching, the terms in a batch combine in different numbers[20] when matched with documents in a best match system like Okapi (cf. 7.2.2). Any part of the user's query not represented in the batches are therefore excluded from the search process. One way of (partially) compensating this is to include the original user input terms in the final search statement (cf. 7.3.3). However, since usually 'within document collection frequency' of thesaurus terms are substantially less compared to most of the free text terms, the weights of thesaurus terms tend to be higher compared to free text ones, generally. As the number of thesaurus terms in a batch also tend to be more than the user input terms, the effect of inclusion of the user's own search terms in the final query statement is likely to be off-set for to a large degree by the thesaural terms.

As a conclusion it can be said that, when one or more of the user's search terms are not represented in the batches, the final search statement is bound to be in most cases partially representative of the user's query, which may result in poor overall retrieval effectiveness.

7.4.2 Generation and ranking of linked terms

One of the most important design decisions related to the generation of batches is the number of levels to be expanded in the spreading activation process (cf. 7.3.2.2).

[19] In some cases although a user input term is not represented by any of the source terms it may appear in the batches as one of the linked terms. When referring to a user term appearing in a batch, it is meant the term itself or its stem matches with at least one of the component terms (or its stem) of a compound thesaural term (cf. 7.1.1 and 7.1.2) as discussed in 7.3.1.4.

[20] Considering that in the experiments performed (cf. 8.3) number of terms in a single batch ranges between 5 to 8, it is rarely the case that there exists a document that is indexed by all the terms contained in a batch. However, it is in general possible that there could be several documents indexed by 3 or more of the terms in a batch.

Most systems that use the spreading activation to find relevant terms from a knowledge-base expand such that, either the maximum number of expansion level allowed is reached, or all source nodes are connected (e.g. Chen, 1992; see also 2.2.2). The maximum number of expansion level is usually limited to 2 on the premise that, terms beyond two levels from a given source term are usually less likely to be relevant to the user (e.g. Chen, 1992; Paice, 1991).

The systems developed in this project (*KBS-1* and *KBS-2*) differ from the others in that, the spreading activation is allowed to progress for up to 4 levels rather than 2 from a source term (maximum of 7 levels altogether from the both ends). The major reason for this as discussed in 7.3.2.1 is that, the main objective of the systems devised is to explore different search areas of possible interest to the user, rather than to find documents just resembling to those described by the original user search terms (cf. 6.3).

One of the reasons for this particular limit of 7 expansion levels chosen is that, in the pilot study of 8.2.1.3 it became apparent that, a batch with 8 thesaurus terms (i.e. 7 links or levels of expansion, involving 7+1 nodes) is about the limit at which the batch as whole stops being a coherent set. Coherence as would be recalled from 7.3.2 is one of the conditions set for the batches to fulfil. The other rather practical nevertheless important reason for this limit is that, computation of the paths beyond 7 levels becomes too expansive in terms of both time and computer memory required. These two considerations, together with the observation that most terms connect in any case at less than 7 levels of expansion helped to determine the maximum level of expansion allowed in the systems.

Another important characteristic of the systems developed here is that, the source nodes *always* connect. More accurately, the two source nodes (terms) are chosen such that they are connected in the semantic network (cf. 7.1.3) in at least 2, at most 7 levels of expansion (cf. heuristics *III.i* in 7.3.1.4). In many other systems, connection of the source nodes is not always an absolute requirement (e.g. Chen, 1992). The number of expansion levels is normally limited by 2, regardless of whether some source terms remain unconnected (cf. 2.2.2). The difference between the approach adopted in this project and the one exemplified by the systems mentioned above can be accounted by a number of differences in the basic design philosophy.

The first major difference as discussed in the previous paragraphs is to do with the sort of terms to be found by means of the spreading activation. As noted above, the approach adopted here aims to find terms which may not be *foreseen* by the user initially, but having been made available, could be taken up by the user subsequently, i.e. could help the user to prescribe new relations or relevance criteria (6.3.3, 6.4.3). This means that, as discussed a few paragraphs earlier, the spreading activation should not be limited to a few immediate levels from the source terms. Instead of 2 levels of expansion from the source terms (as proposed by some of the systems mentioned earlier) the total number of expansion levels between the two source terms could be up to 7 in both *KBS-1* and *KBS-2*, resulting 4 levels of expansion from one of the source terms, and 3 levels from the other. In most cases, this is sufficient to connect any two terms in the semantic net.

However, beside the above discussed reason there are other reasons for specifically requiring connection of the two source terms. The first reason for specifically stipulating that the two source terms must always connect is to do with the fact that, the batches are conceived to function as conceptually coherent sets that describe particular portions of the knowledge space (cf. 7.3.2; 7.3.2.1). Therefore, the approach adopted in this project differs from some the others in that, the task of the spreading activation is to describe different areas of the domain of inquiry, rather than merely to find terms related to the user input search terms. Given the assumptions of semantic network as representation of a given knowledge domain and source terms as the representation of the concept(s) expressed in the user's search statement (cf. 7.1.3 and 7.4.1), it follows as a logical consequence to connect the two source terms to describe the knowledge domain addressed by the user's query.

Another reason for always ensuring the connection of the source terms is to do with the objective of constraining the spreading activation (cf. 7.3.2.2). Unless the spreading activation is constrained by imposing some conditions, it results in too many connections most of which are not useful for any purpose and the process becomes computationally uneconomic (Chen, 1992; Cohen & Kjeldsen, 1987; cf. 7.3.2.2). The approach adopted in this project is to constrain the propagation in the semantic network by making sure that the source terms always connect (cf. 7.3.2.1), rather than constraining the process by imposing other conditions, such as, restricting the number of expansion levels, or by means of other more complicated methods, such as, attaching value or weights to particular links or types of links (cf. 7.3.2.2). In other words, the condition that the two source terms must connect, apart from performing the functions described earlier, acts as a constraint such that, many other possible activation paths in the net which are not expected to be useful to the user are avoided.

Another interesting point to discuss is the effect of activating redundant and a priori determined number of levels in generation of the batches (cf. 7.3.2.2).

Lets assume that at *l* levels of activation there is a batch consisting of the following nodes: N_{b1} N_{b2} ... N_{bl+1}. At one more level of expansion (i.e. *l*+1, one redundant level of expansion) it is possible that, there may exist no batch that contains all elements of this batch. The pertaining question here is the following: if all elements of the above batch are not picked at an extra level of expansion, does this mean that expanding an extra level causes loss of information instead of generating more information?

The purpose of introducing redundancy in expansion is clearly to gain more information regarding the subject domain of inquiry (cf. 7.3.2.1 and 7.3.3). Since the distance between two terms in a semantic net is assumed to be indicative of their conceptual closeness, losing a term which is connected immediately to one of the source terms at the preceding level of expansion might well mean loss of important information.

It is indeed probable that at an extra level of expansion the 'N_{b1} N_{b2} ... N_{bl+1}' chain can be broken, therefore, the information represented by it lost. However, although this is a possible scenario, it is highly unlikely to occur in practice. To lose this chain when an extra level of redundancy is introduced, it is necessary that no common term to exist between any two of the subsequent nodes from N_{b1} to N_{bl+1} in the path illustrated above. As most Inspec thesaurus terms have many connections (cf. 7.1.1) this is relatively an unlikely, although not impossible, event. However, more importantly, it is worth to realize that, the current wisdom in semantic net approach to IR assumes that the number of links (or common nodes/terms) between any two node is indicative of the strength of association between them (Paice, 1991, p. 437; see also 2.2.2). Therefore any broken link at an extra level of activation is likely to be a weak one and probably does not contain quality information.

Hence, as a conclusion it can be said that in the actual activation process, it is highly unlikely that any of the links between the terms in a path would be broken because of an extra level of expansion. In any case, information lost in this way likely to be insignificant and could be neglected without any real damage.

7.4.3 Searching the database

The main problem with query formulation and searching using the method developed in this project is the difficulty of combining different terms inferred from different sources, hence providing different sorts of evidence in relation to the user's query.

In a Boolean system for instance, the two source terms could be expected to be coordinated with the 'AND' relation, as they are assumed to represent two distinct concepts present in the user's query (cf. 7.4.1). However, the relation of the other terms in the batch to these two term and hence to the user's query is not self evident. How to coordinate the intermediate terms in a batch with the source terms, from which they are inferred, therefore is not clear in terms of the Boolean logic.

Were there detailed information in the Inspec thesaurus regarding the specific types of relations between the terms, this might help us to understand the inter-term relationships, therefore their expression in Boolean logic.

Faceted classification schemes embody more information regarding term relationships. However, it is doubtful that information embedded in such schemes would be enough to provide enough intelligence to determine accurately the complex relationships that emerge when several different thesaurus terms are brought together in a single batch.

The problem is amplified when coordination of the thesaurus terms with the original user search terms is desired. There is no established search tactics or strategies in Boolean searching literature to address the issue of combining controlled vocabulary terms with the user's own free text search terms. Terms from a thesaurus normally replace the user input free-text terms in traditional Boolean searching rather than combined with them in someway.

These problems are not resolved in probabilistic best-match systems. In terms of Okapi's probabilistic model, the problem is how to merge terms that are qualitatively different (e.g. free text versus controlled vocabulary), and pertaining to different sorts of evidence (e.g. user input terms versus thesaurus terms inferred from the user input terms, i.e. the source terms, and terms inferred from the source terms, i.e. the linked terms).

Okapi is designed to work with evidence derived *direct* from the user, i.e. user input search terms and relevance feedback information. Therefore, only the first sort of evidence noted above (i.e. user input free-text search terms) can be directly interpreted by Okapi. The other two other sorts of evidence, namely, source terms inferred from user input terms, and linked thesaurus terms that are inferred from the source terms need to be interpreted in a novel way and merged into a consistent whole.

It seems logical for instance that, weights assigned by F4/WPQ formula to the source and linked terms need to be adjusted, perhaps lowered, according the quality (directness) of the evidence. Alternatively, the weights of the user input terms might be increased by a certain factor to reflect the fact that they are derived by direct evidence from the user unlike the source and linked terms which only provide indirect evidence.

Finally, the problem gets even more complicated when the quality ('well-definedness') of the source terms and the terms inferred from them are considered. When there is no exact matching terms in the thesaurus the two source terms are selected from thesaurus terms that do not constitute an exact match, as dictated by heuristic *II.vi* in section 7.3.1.4. In such a case the selected source terms might well be 'under-characterised', that is, may contain as few as one of the user input terms or its stem. This is also true when there is only one exact matching term and the second source term is selected from non-exact matching terms in accordance with the heuristics *II.i* and *II.iii*, which may again be under-characterised. When the exact match is of type *I.iii-b*, it may also be under-characterised as it is possible to make an exact match of this type by one or a few of the user input search terms.

In all the above cases, it may be desirable to lower the value of such under-characterised terms and the terms inferred from them. Again there is no obvious heuristic which suggests how this might be done.

There are no clear cut answers to any of the above noted problems and their investigation cannot be undertaken within the limits of this project. These problematic points are therefore put aside for future work. In the mean time, simpler assumptions regarding these points are made for the purposes of the present project (as discussed throughout the present chapter).

Chapter 8
Evaluation Experiments and Results

In this chapter, first, evaluation methodologies and measures used in IR experiments are reviewed (8.1). In section 8.2, the experimental methodology employed in this project is described. The results of the experiments performed are presented in section 8.3. The last section (8.4) compares and discuss the results of this study with that of the CILKS project.

8.1 Evaluation in IR

In section 8.1.1 various approaches to evaluation in IR work are discussed briefly. In the following section (8.1.2) effectiveness measures used in retrieval experiments are reviewed. In the last section (8.1.3), issues involved in evaluation exercises with real users (which is the approach taken in this project) are briefly discussed.

8.1.1 Experimental approaches

Evaluation in IR is a complex problem that has been a source of innumerable discussion and debate. This complexity arises in part from the large number of variables involved in any real life interaction with a retrieval system.

The boundary of a retrieval system itself is a highly problematic and debatable matter. It is generally agreed that at least the user as an individual subject with some cognitive states should be included in the evaluation process. However, it is possible to extend the boundary of the system beyond the individual user to include organizational, societal and even political levels. In a recent paper (Karamuftuoglu, in press), I have challenged the adequacy of the individual user as the limit of an IRS in light of the advances in network-centric computing and retrieval. In the present study however, I had to contend to draw the boundary of the retrieval system at the level of the (individual) user as in many other studies.

Once the boundaries of a retrieval system are drawn, system effectiveness can be studied at various levels (Bawden, 1990):

- at the level of a component part or sub-system of the IRS
- at the level of the whole IRS

Typically, to study the effects of various contributing factors to the performance of an retrieval system, some component part (or sub-system) of the whole system, such as the matching function or the indexing language is isolated and studied in detail. Some studies take IRS as a whole and aim to evaluate its overall performance. In both approaches, the end user may be included or excluded from the system's boundaries.

IRS evaluation experiments can also be classified in other ways (Robertson & Hancock-Beaulieu, 1992):

- laboratory versus operational
- black box versus diagnostic
- qualitative versus quantitative

Ideally, all retrieval experiments should conform to the real operational conditions. However, the conflict between controllability, repeatability and observability of the experimental variables and the reality of the operational conditions often results in a less than desirable realism in experimental designs. The realism issue has been centred on the query and the relevance judgements since the early days of the IR experiments. Some aspects of the users' requests and relevance judgements can be simulated in *laboratory experiments*, using for instance test collections with a set of requests and a corresponding set of relevant documents, as in the paradigmatic Cranfield tests. However it is extremely difficult and perhaps impossible to simulate realistically a highly interactive operational situation where the users' queries and relevance criteria may change dynamically as a result of the system's responses and other contextual factors. On the other hand, even in *operational experiments* involving real people with real information needs some sort of context control is clearly needed to be able to make some experimental observations and isolate the effects of the experimental variables on the system's performance.

Black box experiments treat the IRS as a black box and aim to observe the relationships between input and output states. In this approach the boundaries of the system should be clearly defined and input and output states must be measurable or observable. Most experiments involving test collections employ black box type approach to evaluation. By performing large number of such experiments diagnostic inferences to improve the system's effectiveness can be drawn. This type of experiment is usually conducted to decide between two or more competing systems or principles. *Diagnostic experiments* on the other hand aim to directly identify categories of failures in terms of system features. This type of experiments are expected to lead directly to recommendations for improving performance by modifying the system's internal mechanisms. Although these two approach have different objectives and employ different experimental methods in practice, parts of both approaches can be found in many experimental designs.

Most traditional IR experiments, for instance the Cranfield tests (Cleverdon, 1967), put emphasis on the *quantitative* measures, such as *recall* and *precision*, of system performance. However as Robertson and Hancock-Beaulieu (1992) note there are several *qualitative* aspects involved in apparently quantitative tests. The initial assessment of the request-document pairs in such experiments as well as a final assessment of which system(s) perform better than other(s) are qualitative in nature. Some experimental designs, such as the OPAC Evaluation project (Hancock-Beaulieu, 1990; Hancock-Beaulieu et al., 1991) take a more explicitly qualitative form and involve qualitative assessment of users' information seeking activities and the users' perceptions of those activities. The qualitative judgements are then usually subject to quantitative analysis cumulated over users.

8.1.2 Measures of effectiveness

The most frequently used measures of retrieval effectiveness are precision and recall. There are also other measures, such as *fallout*, which are similarly based on the contingency table (fig. 8.1) and used extensively. These measures are defined below.

To illustrate recall, precision and related measures consider a system that makes *n* binary decisions, each of which has exactly one correct answer: YES or NO, either a document belongs to a particular query or does not. Recall, precision and other related measures can then be calculated by comparing the system's decision with some 'standard of correctness' usually provided by the expert indexers who manually assigns the documents to the queries. The following contingency table summarizes the relationship between the system's decisions and the expert judgements.

	Yes is Correct (Relevant)	No is Correct (Nonrelevant)	
Decides Yes	a	b	a + b
Decides No	c	d	c + d
	a + c	b + d	a+b+c+d=n

Figure 8.1: Contingency Table for a Set of Binary Decisions

The following measures can be defined in terms of the contingency table:

(1) recall = a/(a+c)
(2) precision = a/(a+b)
(3) fallout = b/(b+d)

In words, recall is the proportion of relevant documents that the system assigns to the query. Precision is the proportion of documents assigned to the query by the system correctly. Fallout is the proportion of nonrelevant documents that the system assign to the query. An ideal system would have recall and precision of 1. Fallout is an alternative to precision. An ideal system would have fallout of 0.

It is possible to define a single measure of effectiveness in terms of the contingency table (Lewis, 1995):

(4) error rate = (b+c)/(a+b+c+d)

Another measure sometimes used in retrieval experiments is overlap:

(5) overlap = a/(a+b+c)

This measure is symmetric with respect to both b and c and so sometimes used to measure how much two retrieval decisions are alike without defining one or the other to be correct (Lewis, 1991).

However, the most frequently used measures are recall and precision (or fallout). When taken together recall and precision provide a useful measure of the system's performance.

There are two methods of calculating average recall and precision for a set of *m* queries and *d* documents, which requires a total of *n=md* decisions are made. *Microaveraging* considers all *md* decisions as a single group and calculates recall, precision as defined above.

Macroaveraging, on the other hand, computes these measures separately for the set of d documents associated with each single query m_i, and then computes the mean of the resulting m effectiveness values (Lewis, 1991).

In text retrieval, macroaveraging has been favoured partly because it gives equal weight to each user query. A microaveraged recall measurement, for instance, would be disproportionately affected by recall performance on queries with large number of relevant documents (Lewis, 1991, p. 313).

One limitation of the measures based on the contingency table is that they do not take into account the possibility that different errors have different costs. Doing so requires a more general decision theoretic model. The contingency table also requires all decisions to be binary.

In equations 1 to 3 above, zero (0) denominators arise when there exist no relevant documents, no nonrelevant documents, or when the system does not retrieve any documents (ibid). All these scenarios are extremely unlikely when microaveraging is used but are quite possible under macroaveraging. In the context of text retrieval, it has been suggested that (e.g. Tague, 1981), 0/0 in the above cases either should be treated as 1.0 or the query should be discarded, although neither of the solutions is entirely satisfactory.

Another problem with recall is that it is very hard to establish the total number of relevant documents in a collection for a given request. There are numerous reasons for this difficulty. Firstly, who is to determine the relevance of a document to a query. As noted in the previous section when laboratory versus operational experiments is discussed, a test collection with fixed document-query pairs is not a realistic assumption to be applied to operational systems used in real contexts. However, it is practically impossible in any large scale collection to have the user go through each and every document to determine the total number of relevant documents. We will return to this problem when the measures used in this project are discussed in the following section.

Recall also does not give an absolute measure of system's effectiveness in retrieving the relevant documents. Consider the case where only 100 relevant documents exist in a collection for a certain query. The recall in this case will be 80% if the retrieval system retrieves 80 of the total of 100 relevant documents. The total number of missed relevant documents will be 20 in this example. Lets assume in another collection we have 1000 relevant documents for the same query. The same system operating at 80% recall will retrieve 800 of the relevant documents and miss 200 of them. Clearly in the second case for a user who would like to do an exhaustive search, the total number of missed documents would be unacceptably high. This affect of the absolute size of the relevant documents on the system's performance is not readily captured in recall.

Many variations of recall/precision, as well as, other measures of retrieval effectiveness have been suggested and used in some cases. However, recall, precision (or fallout) are the most widely accepted and used measures in the IR community. For this reason, discussion of other effectiveness measures will not be attempted here.

8.1.3 Evaluation with real users

As noted earlier, although evaluation exercises with test collections (such as the Cranfield, and more recently TREC experiments) have the advantage of repeatability and controllability which make possible to compare the results across different systems or retrieval methods, they are often criticised for overlooking many operational or contextual factors. As already mentioned, in experiments involving test collections it is very problematic to establish document-request pairs without making simplifying assumptions regarding contextual and operational factors.

In general, it is desirable to have the same person who has put the query to the system in the first place to make the relevance judgements on the retrieved documents. This would provide a more realistic picture of the IR process, as in operational situations documents retrieved by the system will be evaluated by real users with real information needs in most cases.

Relevance judgements by the originators of the queries, however, raise some points that need to be considered. Firstly, in experiments with real users and queries, recall is not readily calculable, as it would not be possible to know the total number of relevant documents for a given query in advance. In addition to this, user's relevance judgements may be affected by factors other than the subject (or content) relationship between a document and a query. This is sometimes articulated as the difference between relevance and pertinence. It is reasonable to assume that user's previous knowledge of the document to be judged or other documents retrieved or seen previously, as well as particular characteristics of the document (such as language, availability, style, narrative structure etc.) may affect the utility or use of that document to the user. A document, thus, may be relevant however not pertinent if the user has already seen the document or the contents of it is covered by some other documents known to the user or for any of the other reasons mentioned above.

In this project, real users with real information needs are used in the evaluation exercises. The reason for this decision is to have a more realistic representation of the actual retrieval situations.

The users are instructed to judge the relevance of documents regardless of whether they have seen them before. This is worded as follows (appendix A): "You should judge the references according to being relevant or not to the subject domain as described by your query statement, regardless of whether you have seen the reference or the document it refers to before". In this way it is hoped that relevance and not pertinence of documents are measured.

In the two evaluation experiments (see 8,2) carried out in this project, precision values are used as measure of effectiveness. As total number of relevant documents are not available, it was not possible to calculate recall. However, this is not considered as a handicap, since recall is a less realistic measure than precision from the general perspective taken in this project as it implies a fixed number of relevant documents for a given query. This assumption ignores the dynamic nature of IR interaction and the relevance judgements.

Since the systems designed in the project produce a ranked list of documents (cf. 7.2 and 7.3), cut-off points (rank positions) at 5, 10, 15 and 20 documents are used in the tables that show the precision values (see appendices N, and O). There are a number of problems with cut-off points however. Firstly, a fixed cut-off point for each query (such as 5 or 10 documents) does not take into account different number of relevant documents in the collection that may apply to a particular query. For instance, if there are just 15 documents that are relevant for a particular query in the collection and if we apply cut-off point of 20 documents for all queries, we would mistakenly calculate 75% precision for this query even if the top 15 documents retrieved were all of the relevant documents in the collection. Similarly, a query with just 50 relevant documents should be treated differently from a query with 500 relevant documents when a fixed cut-off point is applied. Comparison between queries which have different absolute number of relevant documents is problematic when fixed cut-off points are used. For this reason, macroaveraging does not lead to reliable results when applied to rank-based systems with fixed cut-off points. In this project, therefore, all queries are treated as if they form a single query when precision values are computed (appendices N, and O).

One would expect, on average, an inverse relationship between cut-off points and precision values, i.e. greater the cut-off point is, lower the precision. When different systems are compared at a given cut-off point, higher precision indicates better system performance.

Precision values assume binary relevance judgements: either Yes or No. A degree of satisfaction is not reflected in binary judgements (Robertson & Belkin, 1978a). Saracevic (1971) suggests a three-point relevance scale which can be collapsed into two when precision is calculated:

- relevant
- partially relevant
- nonrelevant

A relevant document is a one, which on the basis of its contents (or the information it conveys) is considered to be related to the user's query. A partially relevant document is a one, which on the basis of its contents is considered to be related to the user's query only in some part. A nonrelevant document is a one, which on the basis of its contents is considered to be not at all related to the user's query. This three-point scale can be collapsed into two by considering partially relevant documents either relevant or nonrelevant when precision values are calculated. This approach is used in the two retrieval experiments performed in the project.

A mixture of quantitative and qualitative approaches to evaluation is employed in the experiments carried out. Users' relevance judgements which give a quantitative measure of retrieval effectiveness are supplemented with qualitative assessments of the systems' performance by the users. These experiments and their results are described in detail in sections 8.2 and 8.3 below.

8.2 Experimental design

The experimental design consists of two separate experiments, referred as experiment 1 and experiment 2, which aim to evaluate various design objectives set in chapters 6 and 7 (see in particular 6.3, 6.4 and 7.3) of the corresponding knowledge-based systems, *KBS-1* and *KBS-2* respectively.

The design objectives and consequently the design of the evaluation experiments for both systems are similar, except that in experiment 2 performance of *KBS-2* is compared to the benchmark Okapi retrieval system. In experiment 1, comparison of the performance of the individual batches are done against the sets of terms generated by combining the unique terms taken from the individual batches, instead of the Okapi system (see below sections 8.2.1.2 and 8.2.2.2).

In the following section (8.2.1) the design objectives of experiment 1 and 2 are stated. This is followed by detailed description of the experimental procedures employed in the experiments in section 8.2.2.

8.2.1 The objectives of the evaluation experiments and the pilot study

In the following two sections (8.2.1.1 and 8.2.1.2) the objectives of experiment 1 and 2 are explicated. In 8.2.1.3, the pilot study conducted prior to formal experimentation is described briefly.

Experiment 1 is conducted to test the success of.the design objectives of *KBS-1* set in 6.4.3. Experiment 2 is performed to test the performance of *KBS-2*, the other system developed in this project whose design objectives are similarly set in 6.3.4.

Design objectives of the both systems on the whole are similar except that *KBS-2* aims to achieve higher retrieval effectiveness (cf. 7.3.1.3 and 7.3.2.3) in terms of the traditional measure of precision used in IR research.

8.2.1.1 Experiment 1

The first experiment, experiment 1 is set up and performed to test *KBS-1* in relation to the design objectives established in chapters 6 and 7 (see in particular section 6.4.3). In summary the main design objective of *KBS-1* is to:

- help the user to explore the different parts of the knowledge-base related to the area defined by the user's original search terms

It is assumed that the knowledge base constructed using the Inspec thesaurus (7.1.3) represents subjects conveyed by the documents in the database. An important task of *KBS-1* (like *KBS-2*) is to describe or outline the contents of the database (the subjects it covers) to the user in a concise manner so that the user can infer (abduct) the relationships between different concepts represented in the database. This information can then be used to prescribe new relations between terms (concepts) to generate new ideas.

As discussed in chapters 6 and 7 (see e.g. 6.2.2.2), these new relations or ideas may constitute an invention (moderate or radical) or it may be that they are new to the user although already known publicly.

To achieve the above described objective, the system should:

- present groups or batches of conceptually related terms that describe parts of the knowledge base in such a way that they make sense to the user of the system

In other words, the user of the system should be able to perceive the batches as meaningful wholes that describe specific subject areas. It should also be possible for the user to distinguish between the batches in a meaningful way, i.e. to detect sensible differences between them. In section 8.2.2.1 below the details of the experimental set-up devised to evaluate these functions of the system are described.

8.2.1.2 Experiment 2

Experiment 2 is devised to evaluate *KBS-2*. The design objectives of *KBS-2* are very much similar to *KBS-1* (cf. 7.3.2.3) and cover those discussed in the previous section (8.2.1.1). In addition to the design objectives described above, *KBS-2* aims to optimise the user's original query (cf. 7.3.2.3). In other words, it has the objective of optimising the list of documents generated by each batch.

To evaluate the retrieval effectiveness of *KBS-2*, each list of documents retrieved by the system is compared with the documents retrieved by the standard Okapi system which is taken as a benchmark. Experimental set-up used in the evaluation of the above stated objectives of *KBS-2* is discussed in section 8.2.2.2.

8.2.1.3 The pilot study

Prior to formal experimentation, previous searches of the users of Okapi are analyzed to experiment with the parameters that affect the performance of the KBSs. For this purpose, the logs (7.2.1) of the previous searches performed on the Inspec database (document collection) are examined. The logs contain the original user input search terms, the details of the documents retrieved by the system, the relevance judgements of the users and the terms selected by Okapi for automatic query expansion.

The purpose of the pilot study is to establish the heuristics needed for the knowledge-based systems developed (cf. 7.3). In particular, the effect of the cut-off points (weight humps) and various weighting algorithms on the selection of the source terms and ranking of the batches (cf. 7.3) are investigated.

The approach adopted in the pilot studies is subjective and exploratory. Two sources of information embedded in the logs proved to be particularly useful: the relevance judgements and terms automatically extracted by Okapi for query expansion (7.2.2). Various weight humps and minimum values of *R* and different weighting algorithms (cf. 7.3.1.3) are tried to select both the source terms and the batches. The effect of these parameters are (subjectively) assessed by comparing the documents retrieved by a specific combination of the values of the parameters with the user's relevance judgements for that particular query. Also terms extracted from the relevant documents by Okapi for inclusion in the search statement are examined and compared with the source terms selected and the terms linked to the source terms in the batches for different values of *R*, weight humps, and different weighting formulae. By comparing the terms selected by Okapi and those by the KBS, an impressionistic assessment of the effect of the various parameters and heuristics used in the knowledge-based systems was possible.

8.2.2 The methodology of the experiments

In this section, first, detailed description of the procedure followed in experiments 1 and 2 are given, in 8.2.2.1 and 8.2.2.2 respectively. This involves description of the particular steps of the procedure followed and questions directed to the participants of the experiments.

These two section is followed by description of the selection of the queries and users utilized in the experiments in 8.2.2.3 and the document collection and the retrieval systems in 8.2.2.4.

8.2.2.1 Experiment 1

In experiment 1 the main objective is to evaluate the knowledge-based system designed (*KBS-1*) in terms of its usefulness for suggesting new terms, new areas of search to the users of the system. The following description of the experimental procedure will explain how this is done.

Each user taking part in the evaluation exercise is first presented with a list of thesaurus terms (appendix E) and asked to rank them in the order of decreasing importance for their search. The terms in the list are derived from the batches generated from the original user input search terms. The generation of the batches has been described in section 7.3.2. For the experiment highest ranking ten batches are used. Unique terms taken from these batches are put in a single list in descending alphabetical order.

Each user is instructed first to select all the terms from the list that he/she considers good for the purpose of the search. The user is then instructed to rank them in decreasing order of importance. After this operation, the users are asked if any of the terms in the list represents new ideas for them (appendix E). It is explained when necessary that, in the context of the experiment, new idea means, concept(s), term(s) that the user was (were) not previously aware of or not originally part of the user's intended search.

The above described procedure comprise the first stage of the evaluation exercise. The forms given to the users explaining the procedure involved at this stage, together with the documentation provided for the rest of the experiment is reproduced in appendix A.

The second stage of the experiment consists of the assessment by the users of the batches of terms generated. The top ten batches generated are printed on a separate sheet and shown to the user.

The users are instructed to assign one or more of the categories provided in another sheet to each of the ten batch listed (appendix F). The categories that the users can choose are divided into two broad groups. In the first group there are two categories that mark the two ends of the spectrum. Category A says that the batch as a whole looks good for the user's search purpose. Category B is the polar opposite of category A. It says that none of terms in the batch are good for the user's search purpose.

If the user cannot define a batch by either of the above two categories, six other categories are provided in the second group can be used for describing a batch. These categories are reproduced in appendix A.

The purpose of these additional six categories is to break the binary opposition between the two categories described above and provide finer distinctions between batches. This is hoped to elicit evidence regarding the ability (or lack of it) of the batches to define a subject or a knowledge space as a semantically coherent whole. Of particularly interest from the point of view of the main objective of the experiment are categories E, F and G.

Category F is intended to be used for batches that represent *new* ideas, hence it is worded as following: "The batch contains terms that represent new ideas which are useful to my search".

Category E which reads as "The batch contains terms that are marginally related to my search (or of secondary importance)", and category G which is worded as "The batch contains term(s) that represent ideas which is/are part of the general domain of my search, however not directly useful to me" are included in anticipation that the users are indeed think or able to think in terms of contiguous knowledge domains that can be visualised to constitute parts of a larger knowledge space (cf. 6.3.2).

The other categories are: C ("some of the terms in the batch are good for my search purpose"), D ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement") and H ("None of the above"). The users are allowed to mark a batch with more than one category if they feel that none of the categories alone applies.

The above described two stages of experiment 1 are intended to gather evidence regarding the success or failure of *KBS-1* in helping (or prompting) the user to *prescribe* new connections between concepts and to test whether individual batches can be regarded to represent subject areas or sub-sets of subject areas in a conceptually coherent way (cf. 6.4.3).

In the final stage of the evaluation, users are presented with a randomly ordered list of documents retrieved in response to their query and asked to provide relevance judgements on them.

The list is derived from documents found by each batch combined with the original user free-text search terms as explained in 7.3.3. The top ranked four batches are used for this purpose (see appendix D for the batches). In addition to the top ranking four batches, a fifith barch is constructed by taking all the unique terms from these four batches. This fifth batch is similarly combined with the users original search terms to retrieve documents.

The objective in creating a batch consisting of unique terms from the top ranking four batches is to experiment whether this new batch has a better retrieval effectiveness in terms of relevant documents found (see below and section 8.3.1.4).

Top 20 documents found by all five batches are combined in a single file and presented to the users in a random order after the duplicate documents are removed. The users are then asked to mark each document as "Relevant", "Not-Relevant", or "Partially Relevant" regardless of whether they have seen the document before (see appendix A).

The purpose of this part of the experiment is to see how each batch perform in terms of users' decisions regarding the relevance of the documents retrieved.

The information gathered at this stage of the evaluation process gives an indication of how the rank of the batches as determined by the system compares with the users' ranking of the batches as derived from the user relevance judgements (appendix N). This information is used to compare the ability of various weighting formulae mentioned in section 7.3.2.3 (appendix P). The results from this stage of experiment 1 are used to select the weighting function used to rank the batches in *KBS-2*. This aspect of the experiment is discussed in section 8.3.1.4.

8.2.2.2 Experiment 2

Experiment 2 is performed to evaluate the system referred as *KBS-2*. The design objectives of this system are similar to *KBS-1* except that it also aims to have increased effectiveness (precision) in its output.

The methodology used in experiment 2 is therefore, similar to that used in experiment 1 (8.2.2.1), except that the output of the system is compared to the output of the benchmark Okapi system.

To compare the effectiveness of *KBS-2* with that of Okapi, top twenty documents retrieved by the highest ranking four batches are combined with the top twenty documents retrieved by the Okapi system. The unique documents are picked up from this combined list and presented to the users taking part in the experiment in a random order. Similar to experiment 1, the users are then instructed to make relevance judgements for each document in the list as "Relevant", "Not-Relevant", and "Partially Relevant".

The steps followed in experiment 2 are exactly the same as those followed in experiment 1, and the results of the experiment discussed in sections 8.3.2.

8.2.2.3 The users

As indicated in 7.2.1, search logs of the users of the operational Okapi system are kept automatically and routinely for all registered users of the system. The user logs were examined for the purpose of the evaluation experiments to identify the users that had recently used the system. From these logs, owners of the queries with at least two distinct terms are selected.

The reason for selecting queries with at least two distinct terms is to do with the design objectives and constraints of the knowledge-based systems (*KBS-1* and *KBS-2*) developed in this project. As it is discussed in detail in 7.3, the systems are developed to assist the users whose queries have more than one aspects, or facets in the broad sense of the term. Therefore, for the evaluation of both *KBS-1* and *KBS-2* users whose queries (search statement) appears to contain at least two distinct concepts are used.

It should be noted here that, selection of the users was inevitably subjective to some degree in that, the assessment of the presence of distinct aspects in a query was performed without any formal procedure. In general, as far as a query contains more than one term, and the terms 'looked' not describing exactly the same concept or synonyms of each other, it is accepted as a legitimate query that can be put into the knowledge based systems developed.

Once the appropriate queries and their owners are identified the procedure described in 7.3 for generating the batches and searching of the database was followed.

In total twenty-five queries were re-run on the knowledge based systems in experiments 1 and 2. It was not, unfortunately, possible to convince owners of the all selected queries to participate in the experiments. Consequently, out of total of targeted twenty-five users, sixteen in total eventually completed the experiments. The results of the experiments are discussed in detail in section 8.3 below.

The users are mainly consisted of M.Sc. and research students at computer science, information science/systems, business computing and technology and electrical engineering departments. There are also a few undergraduate students and academic staff involved in research and teaching at the above named departments of City University.

8.2.2.4 The document collection and the retrieval systems

The document collection used in both experiment 1 and 2 is comprised of 314427 documents taken from a part of the Inspec database related to the information technology and computer science fields.

The documents are consists of such bibliographic records as title, author, year and source of publication, abstract, descriptor terms taken from the Inspec thesaurus (cf. 7.1.1) and free-text keywords.

The details of the knowledge based systems used in the experiments and the Okapi system used as a benchmark in experiment 2 are given in sections 7.3 and 7.2, respectively.

8.3 The Results

The results of experiment 1 and 2 are discussed below in sections 8.3.1 and 8.3.2 respectively.

The results can be grouped in two broad categories for both experiments. Some of the results are related to the usefulness of *KBS-1* and *2* in stimulating new ideas for exploration and describing the queries of the users of the systems. The other part of the results is concerned the effectiveness of the systems in finding relevant documents.

Section 8.3.1 and 8.3.2 which deal with the results of experiment 1 and 2, respectively, are further divided into four subsections.

First, general observations regarding the usefulness of the systems for stimulating creativity and the cognitive aspects of the batches used in the retrieval process are presented (8.3.1.1 and 8.3.2.1).

This is followed by the discussion of the results related to the creative/cognitive aspects of the systems in detail for each of the participants in the experiments (8.3.1.2 and 8.3.2.2).

For each experiment, one of the participants is invited to go through a further set of questions/tests to elicit more information regarding the creative/cognitive aspects of the systems. The results of these tests are discussed under the heading of "the diagnostic evaluation" in sections 8.3.1.3 and 8.3.2.3 for *KBS-1* and *KBS-2* respectively.

Finally, the results pertaining to the retrieval effectiveness of the systems are presented and discussed in sections 8.3.1.4 and 8.3.2.4.

8.3.1 Experiment 1

As discussed in section 8.2.2.1 the experimental set-up of experiment 1 consists of three stages. The first two stages aim to elicit information regarding the usefulness of *KBS-1* in suggesting new ideas to the users and the effectiveness of the batches containing linked terms in describing a coherent set of concepts that describe a subject area.

The third stage involves assessing the retrieval effectiveness of the system by having the relevance judgements of the actual users of the system.

In the following sections, results regarding both aspects of the evaluation experiment are discussed. First in section 8.3.1.1, general observations regarding the results of the first two stages of experiment 1 are presented. This is followed by, in section 8.3.1.2, one by one discussion of each of the users' results related to the stages one and two of the experiment. Section 8.3.1.3 presents detailed analysis of the results of one of the participants related to the first two stages of the experiment. In the final section (8.3.1.4) the results pertaining to the retrieval effectiveness of *KBS-1* are presented and discussed.

8.3.1.1 General Observations

The primary objective of *KBS-1* is to stimulate the users to explore different areas of the knowledge space (document collection) by exploring the relations between concepts (thesaural terms) in the knowledge base.

To evaluate the effectiveness of this aspect of the system, two sets of questions are designed. The first set of questions instructs the user to select and rank terms that he/she considers useful from a list of terms derived from the Inspec thesaurus. The user, then, was asked whether any of the selected terms represent new ideas for her or him, as discussed in 8.2.2.1. The second set of questions instructs the user to assign one or more of the eight categories to each of the ten batches generated by the system (cf. 8.2.2.1).

Out of eight users that completed this exercise, six of them indicated that either some of the terms given in a single list or some of the batches containing these terms represent ideas that they were not thought/aware of previously (cf. appendices E, and F).

Three of the eight users found one or more of the batches as representing new ideas to them (appendix F). Similarly, four of the eight users indicated that one or more of the terms presented in a single list represent new ideas. Only one user, marked both some of the terms presented in a single list and some of the batches as representing new ideas.

This preliminary observation suggests that users do in fact change their query when exposed to new terms which represent ideas/concepts that either they were not previously familiar with or did not articulate/think prior to the instant they were shown such terms (or batches containing such terms).

The other aspect of this stage of the evaluation exercise is to elicit evidence regarding the semantic/cognitive coherence of the batches generated (cf. 8.2.1.1, 6.4.3).

One can only hope to elicit indirect evidence for this, as any direct question regarding this aspect would be arduous to express in a way that ordinary users of the system could apprehend unambiguously and correctly.

The way to go about eliciting indirect evidence in this experiment is to provide to the users a number distinct categories to assign to the batches. The assumption behind this method is that, if the users are able to choose between different categories and to assign them to the batches (without vocally complaining!), this should be considered as an evidence of the users' ability to distinguish between the batches which can be taken as an evidence of semantic/cognitive cohesion of the linked terms presented in the batches.

Furthermore, as mentioned in sections 8.2.2.1 and 8.2.2.2, one user in each of the experiments is persuaded to take part in a longer, more detailed, 'diagnostic' experiment. The results of this exercise which are discussed in 8.3.1.3 and 8.3.2.3 suggest that the users indeed perceived the batches as consistent semantic wholes that represent specific subjects. Consequently, there were no expressed difficulty in distinguishing between the batches and selecting one or more of the pre-established categories to describe them.

All but one of the users took part in experiment 1 used three or more categories to mark the batches. One user used two categories ("A" and "C").

There are also evidence suggesting that (see 8.3.1.2; 8.3.1.3; 8.3.2.2; 8.3.2.3) presentation of terms as semantically linked concepts in clusters (batches) made up of 6-8 thesaural terms seems to provide enough contextual information to disambiguate the meanings of terms, thus preventing wrong interpretations and associations.

8.3.1.2 Stimulation of creativity and cognitive aspects - individual users

In this section, the results of the first two stages of experiment 1 for each participant are presented and discussed.

a. user/query 1: "expert systems education"

This user selected nine terms from a list of twenty. Of these nine terms three of them are marked as representing new ideas: "teaching", "user interfaces", "user modelling".

The selected terms in user's order of importance are: "intelligent tutoring systems", "expert systems", "education" and "explanation" (both at rank three), "educational computing", "knowledge based systems", "teaching", "user modelling", "user interfaces".

Two of the batches are marked as representing new ideas (category F) by this user. The weight given by *KBS-1* to one of the batches marked as representing new ideas ranks it in the fourth place (Batch 4 in appendix F). The second batch marked as category F (Batch 9) was ranked in the tenth place by the system and therefore was not actually used in searching the database.

Batches 27 and 26 are marked as category A ("looks good as a whole for my search purpose") by the user, which are ranked in the third and eight places, respectively, by the system.

Batch 3 (rank 9) is marked as category C ("some of the terms in the batch are good for my search purpose") by the user. The remaining five of the ten batches are all marked as category D ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement") by the user.

As detailed analysis of the results of this query is given in 8.3.1.3, it will not be discussed further in the present section.

b. user/query 2: "tracking noise edge"

This user selected eight terms from a list of nineteen, none of which has been marked as representing new ideas.

User's ranking of the terms in the order of relative importance are: "tracking", "pattern recognition", and "edge detection", "feature extraction" (the last two are both ranked in the third place), "computer vision", "image processing", "image recognition", "video signals" (the last four share the same rank).

Out of ten batches, two of them are marked as representing new ideas ("F") by the user (Batches 10 and 7, ranks 3 and 4 respectively). The user indicated upon further inquiry that the terms "array signal processing" and the broader term for it the inspec thesaurus "signal processing" represents new ideas for him.

The highest ranked two batch (Batches 16 and 26, ranks 1 and 2 respectively) are marked as categories C, and D. All the others are marked as category F and one or more of the following categories: C, D, E, G. The user indicated that although they contain the terms "array signal processing" and "signal processing" which represent new ideas, they also contain terms that he would not like to include in his search, as they represent ideas not directly related or of marginal importance to his query.

It is interesting to note that although the user did not mark any term as representing a new idea from the list of terms presented, he did so when the terms are shown in the context of the related terms in the batches. This is, of course, the original intention behind the idea of presenting to the users small number of semantically related terms, referred as a batch in this project.

c. query/user 3: "hypertext technical manual database"

The user of this query selected eight terms from a list of twenty-five. Three of these are marked as representing new ideas: "electronic publishing", "multimedia systems", and "technical support services".

The user chose the following terms as most important (in descending order): "hypermedia", "multimedia systems", "user manuals", "system documentation", "technical support services", "electronic publishing", "object-oriented databases", "databases".

Interestingly, none of the batches are marked as containing new ideas. All but one are marked as C ("some of the terms in the batch are good for my search purpose"). The user indicated that, broad terms such as "programming", "programming languages", "systems analysis" seem too general to be useful in his search.

Only, one of the batches, Batch 3, which is ranked in the first place by the system, is marked as category A ("the batch as a whole looks good for my search purpose") by the user.

d. user/query 4: "conceptual graphical query language"

This user selected six terms from a list of nineteen. Two of the six selected terms are marked as representing new ideas: "SQL", "relational databases".

The selected terms ranked by the user in the following order of importance: "query languages", "visual languages", "relational databases", "SQL", "graphical user interfaces", "user interfaces".

None of the batches are marked as representing new ideas by the user. One batch (Batch 57, rank 9) is marked as E ("the batch contains terms are marginally related to my search"). Batch 16 (rank 10) is marked as D ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement"). The user indicated that term "programming" in this batch is too general for his search. Batch 14 (rank 2) is marked as C. The user remarked that "geographic information systems" in this batch seems not useful for his search.

All the other batches in the list, including those ranked at the first, third and fourth places by the system, are marked as category A.

e. user/query 5: "texture detection fractals"

This user selected four terms from a list of fifteen. None of which is marked as representing new ideas.

The selected terms are ranked in the following order of importance by the user: "image texture", "fractals", "feature extraction" and "pattern recognition" (the last two shares the same rank).

Batches 52 (rank 1), 51 (rank 5), 48 (rank 6), 46 (rank 7) are marked as category A ("the batch as a whole looks good for my search purpose") by the user. Batches 49 (rank 2) and 41 (rank 3) are marked as category C ("some of the terms in the batch are good for my search purpose"). Batch 50 (rank 10) is marked as category H ("None of the above"), and the user explained that this batch contains some marginally related terms as well as good ones but also contains two terms ("computer peripheral equipment" and "computer graphics") that he would like to exclude from the search statement. Note that "computer graphics" is also present in Batch 46 (rank 7) and when quizzed about this the user stated that, in the context of the other terms in Batch 46 "computer graphics" does not look particularly threatening, whereas in the other batch (Batch 50) it looks totally out of place (unnecessary).

All the other batches in the list, including Batch 40 (rank 4) are marked as category D by the user.

f. user/query 6: "online information and evaluation of quality and reliability"

The user selected six terms from a list of twenty-six terms. Four of the selected terms are marked as representing new ideas by the user. These are as follows: "information science", "DP management", "management information systems", and "public information systems". The six user selected terms in user's order of decreasing importance are: "information science", "information services", "CD-ROMs", "DP management", "management information systems", "public information systems".

It is worth noting that, the user did not choose either of the source terms used in the generation of the batches from this list. When quizzed about this, he explained that "software reliability" is not what he is after, and although "information retrieval systems evaluation" seems appropriate, his main interest is the evaluation of the contents of the online databases rather than the retrieval systems as such.

All but one of the batches are marked as D ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement") and E ("the batch contains terms that are marginally related to my search (or of secondary importance"). The user explained that terms such as "computer installation", "management", "software engineering" are not directly useful to him, therefore they are marked as "D". The user commented that he also marked the batches as "E" because overall they seem to be related to his main search interests.

Batch 14 (rank 9) is marked as "A" ("the batch as a whole looks good for my search purpose") by the user. As from the initial list of individual thesaurus terms, the user did not choose either of the source terms, he is quizzed this time about why Batch 14 is evaluated as category A. The user replied that although he has reservations for individual terms in the batch, overall it seems to contain useful terms for his search purposes.

g. user/query 7: "polarisation mqw splitter"

This user selected only one term ("polarisation") from a list that contains twenty terms.

It is worth noting here that, the user did not choose one of the source terms, "optical elements" from this list. "Optical elements" is the preferred term for "beam splitters (optical)" and "optical beam splitters" in the Inspec thesaurus which match with the user input term "splitter". When quizzed about this particular decision, the user stated that while she thought optical elements is related to her query, she suspected that it was not exactly what she had in might and this term seemed too general to be useful to her.

Batch 91 (rank 1) and 57 (rank 8) are marked as category B ("none of the terms in the batch are good for my search purpose") by the user. Note that these batches like all the others do contain "polarisation", one of the two source terms which was chosen by the user from the initial list of thesaurus terms shown to her. When quizzed about this, the user indicated that no other term in these two batches are good for her search and this is why she marked them as "B".

Batch 2 (rank 6) is marked as A (""the batch as a whole looks good for my search purpose"). All the other batches in the list are marked as C ("some of the terms in the batch are good for my search purpose").

After the examination of the batches the user once more quizzed about the usefulness of the term "optical elements". Her thoughts did not seem to change as a result of seeing the term in the context of the others that comprise the batches.

h. user/query 8: "cd-rom networking"

The user selected four terms from a list of eighteen. None of the terms selected are indicated as representing new ideas. The four selected terms in the order of the user's ranking are as follows: "network operating systems", "CD-ROMs", "computer networks", and "optical publishing".

One of the batches, Batch 42 (rank 9) is marked by the user as representing new ideas ("F"). Batch 12 (rank 4) is marked as category E ("the batch contains terms are marginally related to my search"). Batches 41 (rank 1), 33 (rank 7), 28 (rank 8), and 21 (rank 10) are marked as category C. All other batches in the list are marked as D ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement").

The overall 'impression' from the above portrayed responses is that, the users indeed found helpful the linked terms organized in batches that display the terms as semantically related aggregates. Batches seem to give a useful indication of the conceptual relations present in the database and the users were able to perceive such relations. There is also some evidence that this sort of relations were useful in explicating previously unknown or implicit relations in the users' query. It can be hypothesised that tracing of relations between the thesaurus terms could even lead to explication of previously non-existent or (publicly) unknown relations between concepts. This could then be used by the users to prescribe new connections and new relevance criteria which constitutes the first step towards making an 'invention', as discussed in 6.2.1.2.

These conclusions seem to be justified by the detailed analyses of the responses of one participants from each experiment which are discussed in section 8.3.1.3 and 8.3.2.3 below.

8.3.1.3 The diagnostic evaluation

The purpose of the diagnostic evaluation is to have an in-depth analysis of the functionality of *KBS-1* related to the first two stages of experiment 1.

Most users, understandably, were not willing to go through a longer analysis of their results. Many of them expressed that going through the three stages described in the previous section was already a highly demanding task. Therefore, it was decided that only one user in each experiment will be approached to recruit for a diagnostic evaluation.

The owner of the first query ("expert systems education") discussed in the preceding section was kind enough to agree to spend extra half hour or so to provide a further analyses of her results.

The first part of this exercise consists of presenting to the user five separate sets of 20 documents each of them containing the highest ranked 20 documents found by one of the five batches used in the third stage of the evaluation process described in 8.2.2.1. The user is then asked to go through each set of documents briefly and to make a speculation regarding which of the ten batches shown earlier might have possibly been used in generating the five sets of documents (the batch list provided again).

The purpose of this test is to find out whether the user could correlate a batch of documents with a batch of terms successfully. This should give some indication about the usefulness of the linked terms in predicting the outcome of the search process, thus their ability in defining conceptually coherent (comprehensible) wholes.

Clearly, the outcome of the retrieval process is not only a function of the terms used in the search process but also the matching function itself. Since the search engine used in this project is a best match one (cf. 7.2), documents matching only a few of the terms in a batch are likely to be included in the top 20 documents retrieved. As many terms are common amongst the batches (in any case, the first and last terms in all batches are always the same for a given query, cf. 7.3.2), it is highly likely that in many cases there could be a considerable amount of overlap between documents retrieved by different batches.

Although the above mentioned factors make virtually impossible to establish a unequivocal relation between a batch and a set of documents, it should nevertheless be possible to some extent to predict what sort of documents are likely to be retrieved by a given batch of terms. This test, therefore, aims to elicit information about the usefulness of batches in representing coherent semantic units of meaning that retrieve documents dealing with same or closely related subjects.

The user in this test indicated that, two of the five sets of documents seemed to be generated by Batch 6, two of them by Batch 27 and could not make a decision regarding the source of the fifth set. She was able to identify the sets of documents retrieved by Batches 6 (rank 2) and 27 (rank 3) correctly. The user failed to correctly match Batch 5 (rank 1) and Batch 4 (rank 4) with the corresponding set of documents. The set retrieved by Batch 5 was matched with Batch 6 and Batch 4 with 27. It should be noted that Batches 5 and 6 differ only in one term (out of total of six). Similarly, Batches 27 and 4 differ in two terms out of six, consequently large number of documents are common in each pair of sets.

More interestingly, the user expressed that she could not identify the source of the fifth set. This set was generated by taking unique terms from all four batches mentioned above (i.e. Batches 5, 6, 27, 4), as discussed in 8.2.2.1. It is therefore reasonable to speculate that, the fact that the fifth set was retrieved by a list of terms that are conceptually less coherent compared to the other four batches used in the experiment, this resulted in a heterogeneous set of documents with varying subject matter and for this reason the user was not able to identify the source of the set among the 10 batches shown to the user.

The next step in the diagnostic evaluation exercise was to ask the user to group similar batches together and comment about the relationships between the batches in the same group. The objective of this test is to elicit further evidence about the ability of the batches to define distinct subject areas. In other words, it was assumed that if the user can detect relations between the batches, this should indicate that the user conceive the batches as conceptually coherent wholes.

The user put Batch 5 (rank 1) and Batch 6 (rank 2) in one group as representing related ideas (about the same/similar subject). Batches 27 (rank 3) and 26 (rank 8) grouped as one. Batches 10 (rank 6) and 11 (rank 7) put in the same group. Batches 4 (rank 4) and 3 (rank 9) formed another group. The user indicated Batch 1 (rank 5) can be put in the last group, but it is probably not so strongly related to the other two batches in the group. Finally, the only remaining batch in the list, Batch 9 (rank 10) was stated not to be related to any other batch in the list. It will be remembered that this was the only batch other than Batch 4 in the list marked as representing new ideas (category F) previously by the user (8.3.1.2). Also, it may be worth noting that four of its six terms (i.e. all the terms except the source terms) are unique to this batch.

It seems from this test that the user was indeed able to identify semantic relations between the batches and distinguish shades of meaning among the batches put in different groups. This finding seems to support the assumption that batches are helpful in describing distinct/distinguishable subjects related to the user's query.

Quizzed about Batches 4 and 9 which the user marked as representing new ideas, the user commented that Batch 4, by bringing to her attention the relation between "social sciences computing" and the other terms in the batch (in particular the terms "expert systems" and "computer aided instruction"), inspired her to explore the "applications of expert systems/computer aided instruction technology in the context of social sciences education". Similarly, the user expressed that Batch 9 contains terms "human factors", "user interfaces", "user modelling" and "explanation" that she did not think of in the context of expert systems and education previously. The user indicated that this batch prompted her to explore the "user modelling/human interface related issues in the context of expert systems used for educational purposes".

When quizzed about which term(s) she did not want in her search in the batches marked as category D, the user replied that she was not interested in the administrative applications of computer systems, consequently she thought terms "administrative data processing" and "educational administrative data processing" should not be included in the search statement.

Commenting on Batch 3 which she marked as category C, the user indicated that the batch on the whole look good for her serach purposes and contains the term "teaching" which she did not think of before and thought that it could be useful for her search purpose. On the other hand, she indicated that the term "social sciences" seems too general, this was why she assigned the category C for this batch.

The overriding impression of this exercise is that, the user did in fact thought that the batches are useful for representing relations between concepts. This seems to support the claim that, this sort of representation can be used with some merit to represent subjects contained in a document collection, hence help the users to explore new areas and prescribe new relations (relevance criteria). It also became evident that, although there is a high degree of overlap of terms between the batches, it was still possible for the user to differentiate between them.

It could be suggested that, it should be desirable to generate disjoint batches that have no common terms amongst them (except for the two source terms). However, there is no evidence that this in particular constituted a difficulty in cognitive/semantic terms for the user. The user in the first part of the diagnostic experiment was not able, as a matter of fact, to correctly identify some of the batches that retrieved the corresponding document sets. This should be, in part, a consequence of the high number of common terms among the batches. However, it is probably more accurate to suggest that the partial matching function used in the search process has more to do with this result than any other single factor.

8.3.1.4 Retrieval effectiveness

KBS-1 has the design objective of presenting to the users batches of terms representing subject areas contained in the knowledge-base that are contiguous to the subject area(s) defined by the user's original search terms. This design objective has been evaluated in experiment 1 and the results of the experiment are presented in the preceding sections 8.3.1.1 to 8.3.1.3.

A corollary to the above main design objective is to optimise the ranking of the documents retrieved by the individual batches (cf. 6.4.3, 7.3.2.3). The second part of experiment 1 is devised to elicit information regarding this aspect of KBS-1. Each of the eight users took part in the experiment were presented with a randomly ordered list of documents derived by combining unique documents taken from each of the five batches used in the experiment. As described in section 8.2.2.1, the highest ranking 20 documents retrieved by each batch are taken and duplicates are removed before shown to the users to obtain their relevance judgements. Four of the five batches used are the top four ranking batches generated by the KC component of the KBS-1 (cf. 7.3.2). The fifth batch is produced by combining the unique terms taken from these four batches. The rationale for the fifth batch is to test whether the unique terms from top ranking batches perform better or worse compared to the individual batches that comprise subsets of the fifth batch. All five batches are combined with the user input free-text terms to search the database as described in 7.3.3.

The results of this part of the experiment are presented in appendix N. In tables 1 to 5 in appendix N, for each batch from 1 to 5, documents that are marked as relevant (R), partially relevant (P), and not-relevant (N) are given for each of the eight users completed the experiment.

The results clearly show that, Batch 5 (B5) performs considerably worse than any of the other four batches (B1, B2, B3, B4). This result seems to suggest that batches of conceptually related terms generated by the KC (7.3) perform better than a list that contains larger number of terms that are not necessarily linked in the knowledge-base.

Another supplementary objective of experiment 1 is to gather information regarding the relative ability of various formulae in ranking the batches generated by the KBS-1. In KBS-1, as it would be remembered from 7.3.2.3, F4 is used in ranking the batches. The user relevance judgements information is used to compare F4 with the other candidate formulae mentioned in 7.3.1.3. These are the association measures used in LEXIQUEST (referred to hereby as Simple; cf. 2.2.1), AID (referred hereby as Doszkocs; cf. 2.2.1), and WPQ (cf. 7.2.2) which is used in Okapi and in the selection of the source terms (7.3.1.3). The objective of this part of experiment 1 is to compare the ranking effectiveness of the above named four formulae, and make use of this information in choosing a formula to rank the batches in KBS-2. As in KBS-2 a major design objective is to have high retrieval effectiveness for each individual batch (cf. 7.3.2.3), the formulae which demonstrates higher effectiveness in experiment 1 is selected to rank the batches in KBS-2.

Tables 1 to 8 in appendix P show the relative ranks of the four highest ranked batches by F4 for the eight users took part in experiment 1, by the above mentioned formulae. These four batches as it would be remembered from the preceding section are used in the experiment to retrieve the documents. The first column in these tables gives the rank of the four batches as a result of the user's relevance judgements. Documents marked as partially relevant documents (P) are counted as relevant in these tables. The second column gives the ranking of the batches as predicted by F4, the third column by WPQ, the fourth by Doszkocs, and the fifth by Simple. In the calculation of the WPQ weights, various parameters involved are assigned the values discussed in section 7.3.1.3.

The tables in appendix P suggest that, in most cases WPQ perform better in predicting the batch ranks than the other three formulae, in particular, in user 1, user 2, user 4 and user 6. It is perhaps worth noting here that, as it would be remembered from 8.3.1.2, user 6 marked only one of the batches, Batch no.14, as category A ("the batch as a whole looks good for my search purpose"). This batch is ranked at the ninth place by F4, second place by WPQ, and first place by Simple (Doszkocs ranks at 15). In user 8, F4 seems to predict better than the others. Doszkocs, and Simple in general seem to be performing worse in comparison both to WPQ and F4 in most of the cases.

The above observations suggest that WPQ is better in ranking batches (in terms of the precision measure used) in comparison to the other formulae tested, and therefore used in KBS-2 in order to achieve higher retrieval effectiveness (cf. 7.3.1.3 and 7.3.2.3). Although, there is not enough data to make statistically significant comparison of the formulae, the impressionistic conclusion drawn above seems to be fairly realistic.

In tables one to sixteen in appendix G, relative positions of the documents marked as relevant and partially relevant by the eight users are shown separately for each of the five batches used in the experiment[1]. The purpose of the information displayed in these tables is to give an indication of whether or not there is a substantial difference between the batches in ranking the relevant and partially relevant documents for a given query. In general, the information presented

[1] Top ranking 500 documents retrieved by each of the five batches for each of the eight users are used in these tables.

in these tables suggest that, substantial number of relevant and partially relevant documents are ranked at noticeably different positions by different batches, although there is a substantial number of documents ranked at similar positions by all or many of the batches. Therefore it may be concluded from this observation that, for a given query different batches tend to have different relevant and partially relevant documents in the top portions (top 20) of their list, in general.

This may be a useful function of the batches in practice, as this characteristic can be used advantageously by the users' in an interactive retrieval situation to explore different areas of the retrieved set. As it is construed that each batch represent a subject area different from the others (and evidence presented in 8.3.1.1 to 8.3.1.3 seems to support this assumption), the users can choose different batches to retrieve relevant documents that address to different aspects of their query.

Finally, KBS-1 seems to be performing worse for some users/queries than the others as can be observed in the tables 1 to 5 of appendix N. One of the main factors that affect the performance of the system seems to be the source terms that are used in the generation of the batches. The results of the experiment presented in sections 8.3.1.1 to 8.3.1.4 above suggests that when there is no good match between the original user input terms and the selected thesaurus terms (the source terms), the retrieval effectiveness deteriorate sharply. For more discussion of this point, see section 8.4 below.

8.3.2 Experiment 2

Similar to the analysis of experiment 1 in 8.3.1, experiment 2 is analyzed under four separate headings below.

In the following section (8.3.2.1), general observations regarding the stimulation of creativity and cognitive aspects of *KBS-2* are presented. In 8.3.2.2 the results regarding the creative/cognitive aspects of the system are discussed for each of the participants in the experiment in detail. Section 8.3.2.3 presents detailed analysis of the creative/cognitive aspects of *KBS-2* for one of the participants in the experiment. Finally, in 8.3.2.4, results pertaining to the retrieval effectiveness of *KBS-2* are dealt with.

8.3.2.1 General Observations

The main objective of *KBS-2* like *KBS-1* is to stimulate the user to explore different areas of the knowledge space (document collection) by exploring the relationships between the thesaurus terms encoded in the knowledge-base of the system.

As in experiment 1 (cf. 8.3.1.1), two sets of questions are presented to the participants to evaluate this aspect of the system. It will be recalled from section 8.3.1.1 that the first set of questions instructs the user to select and rank terms that he/she considers useful from a list of terms derived from the Inspec thesaurus (appendix K). The user is then asked, whether any of the selected terms represent new ideas for her or him (appendix K). The second set of questions instructs the user to assign one or more of the eight categories to each of the top ranking ten batches generated by the system (appendix L, cf. 8.2.2.1 and 8.2.2.2).

Altogether eight users took part in this experiment. Three of the eight participants have indicated that one or more of the batches presented represent new ideas to them (Category F, see appendix L, and 8.2.2.1). None of the selected terms are indicated as representing new ideas by any of the participants in the exercise.

The findings in general seem to support the observation made in experiment 1 (8.3.1.1) that, users do in fact change their queries when exposed to document surrogates (linked thesaurus terms or batches in this case). It can be further hypothesised that, conceptually related terms presented in batches are particularly effective in helping the users to explore new ideas. There is some evidence from both experiments that (8.3.1.2; 8.3.1.3; 8.3.2.2 and 8.3.2.3) users found batches more useful in representing concepts than a list of separate terms.

Since there are large number of common terms between batches that marked as "F" (representing new ideas) and those marked as "A" ("the batch as a whole looks good for the users search purpose") or C ("some of the terms in the batch are good for my search purpose")[2], it may be possible to construe that batches marked as "F" represent concepts related to the user's original query but previously were not known to the user. Detailed analyses of one of the user's responds in each experiment (see 8.3.1.3 and 8.3.2.3) seem to justify this hypothesis.

The other main objective of experiment 2, like of experiment 1, is to elicit evidence regarding the semantic/cognitive coherence or consistency of the linked terms. As discussed in 8.3.1.1, to elicit evidence regarding this aspect of the batches, the users are asked to assign one or more of the eight categories provided to each of the ten batches generated in response to their search terms (appendix L). It is argued in 8.2.2.1 that, the user's ability to carry out this task without overt signs of difficulty would give an indication of the batches ability to represent distinct ideas or subjects.

Seven of the eight users that took part in experiment used two or more of the categories provided to mark the batches[3]. Only one of the users marked all batches with a single category ("C"). This finding seems to justify the conclusion drawn in sections 8.3.1.1 and 8.3.1.2 that, the users indeed found easy to distinguish conceptually between the batches and perceive them as meaningful wholes. Further evidence for this comes from the detailed evaluation exercise performed with one of the participants in experiment 2 (see 8.3.2.3).

8.3.2.2 Stimulation of creativity and cognitive aspects - individual users

In this section, the results of the first two stages of the experiment 2, i.e. selection and ranking of terms and description of batches in terms of the eight categories, are discussed for each of the participants.

a. user/query 1: "object database benchmarks"

The user selected three terms from a list of twenty-four terms. These are the following in the user's order of importance: "performance evaluation", "object-oriented databases", and "database

[2] As a matter of fact, batches that were marked as F and those that were not quite often differ in just one or two terms.

[3] Six of them used at least three categories, one used only two category ("C" and "D") to distinguish between the batches.

management systems".

None of the batches are marked as representing new ideas. Three of the ten batches, Batches 0 (rank 1), 4 (rank 8), and 2 (rank 10) are marked as category "C" ("some of the terms in the batch are good for my search purpose"). All the others are marked as "D" ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement").

b. user/query 2: "computer vision detection"

The user selected three terms from a list of eighteen. These are "edge detection", "feature extraction", and "pattern recognition". The user found all of them equally important for his search purpose, therefore, did not produce a separate ranked list.

It is worth noting that "computer vision" which is one of the source terms used in the generation of the batches was not selected by the user from the list of eighteen thesaurus terms. When quizzed about this, the user expressed that he thought this term seemed too general for his search purpose.

All ten batches are marked as category C ("some of the terms in the batch are good for my search purpose") by the user.

c. user/query 3: "hypertext internet"

This user selected four terms from a list of seventeen. These are in the user's order of importance: "data communication systems", "hypermedia". "multimedia systems", and "visual databases".

One of the batches, Batch 5 (rank 6) is marked as representing new ideas ("F") by the user. Batch 4 (rank 5) is marked as category E ("the batch contains terms are marginally related to my search (or of secondary importance)").

Batch 7 (rank 3) is marked as "G" ("The batch contains term(s) that represent ideas which is/are part of the general domain of my search, however not directly useful to me").

Batches 6 (rank 2) and 8 (rank 1) are marked as "C" ("some of the terms in the batch are good for my search purpose"). Only Batch 10 (rank 10) is marked by the user as category A ("the batch as a whole looks good for the users search purpose").

All other batches are marked as category D ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement").

It is worthwhile to note that the user did not choose the term "computer networks", one of the two source terms used in the generation of the batches, from the initial list of seventeen. This term is one of the two preferred terms for the user input term "internet" in the Inspec database[4].

The results of this user is discussed in detail under "the diagnostic evaluation" heading in section

[4] The other preferred term is: "internetworking".

8.3.2.3, therefore, it will not be analyzed further here.

d. user/query 4: "uv spectroscopy database"

The user selected eight terms from the list of nineteen individual thesaurus terms. The user did not provide a ranking of the terms chosen and indicated that they are all of same importance to his search. The eight terms the user selected are as follows: "chemistry computing", "computerised spectroscopy", "computerised instrumentation", "modulation spectroscopy", "spectra", "spectroscopy", "spectroscopy applications of computers", "time resolved spectroscopy".

The user did not select either of the two source terms from this list. When quizzed about this, the user replied that, both of the source terms "spectroscopy computing" and "infrared spectroscopy" seemed too general for his search purposes.

The user marked Batches 10 (rank 1) and 12 (rank 2) as "A" ("the batch as a whole looks good for the users search purpose"). When the user was reminded that he did not select the two of the source terms from the initial list, he responded that although the source terms are too general, both of the batches seem to be useful as a whole for his search purposes.

Batch 11 (rank 3) and 5 (rank 10) are marked as "C" ("some of the terms in the batch are good for my search purpose"). Batch 9 (rank 4) is marked as "D" ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement"). Batch 4 (rank 5) as "G" ("The batch contains term(s) that represent ideas which is/are part of the general domain of my search, however not directly useful to me"). All the other batches are marked as category "G" and "C".

e. user/query 5: "impact information technology disabled"

This user selected six terms out of nineteen presented in an alphabetically ordered list. These are in the user's decreasing order of importance are: "handicapped aids", "teleconferencing", "telephony" (the last two shares the same rank), "integrated software" and "word processing" (these two also ranked at the same place by the user), and "computer applications".

Two of the ten batches, Batches 6 and 40 (ranks 6 and 7) are marked by the user as representing new ideas ("F"). These two batches are also marked as category E ("the batch contains terms that are marginally related to my search (or of secondary importance").

Batch 6 contains terms "CAD/CAM" and "manufacturing data processing" which the user thought could be useful for her search. The user expressed that these terms when considered with the other terms such as "handicapped aids" and "office automation" in the batch, indicated new areas of application of information technology for disabled persons which she did not think of previously. The user, however, felt that she was not sure whether this course of inquiry would be really useful to her in the end and therefore, marked Batch 6 also as category E.

Similarly in Batch 40, terms "military computing" and "logistics data processing" indicated to the user possible new areas of application of the information technology for the disabled that she did not think of before. However, as the user was not sure whether these terms are really useful to her, she marked Batch 40 as both category "F" and "E".

It is worthwhile to note here that the user in fact did not chose the term "office automation", which is one of the source terms used in generating the batches from the list of nineteen terms initially shown to her. "Office automation" is one of the five preferred terms for "information technology" in the Inspec thesaurus. The others are: "computer applications", "digital computers", "factory automation", and "telecommunication". When the user quizzed about this, she stated that her query is about application of "information technology for disabled people" and not in particular about "office automation".

After the user examined the batches, she was quizzed whether she thought "office automation" is a useful term for her after seeing it in the context of related terms. She expressed that although she was not sure whether this term exactly corresponds to "information technology", it seemed to be a reasonable term in the context of the other terms in the batches. The user as a matter of fact marked Batches 13 (rank 1), 14 (rank 2), 10 (rank 3), 12 (rank 5), 8 (rank 9) and 2 (rank 10) as category A ("the batch as a whole looks good for the users search purpose") which supports the above account by the user.

The above evidence seems to justify the assumption regarding the usefulness of batches in disambiguating the terms, which may have more than one sense or the user is not familiar with, by providing a context.

Batch 11 (rank 4) is the only batch marked as category C ("some of the terms in the batch are good for my search purpose") by the user. Batch 15 (rank 8) marked as "E" and "D".

f. user/query 6: "neural water pollution"

This user selected four terms from a list of fifteen. These are in the user's order of importance: "water pollution detection and control", "neural nets", "water treatment", "artificial intelligence".

None of the batches are marked as representing new ideas. One of the Batches, Batch 12 (rank 4) is marked as category A ("the batch as a whole looks good for the users search purpose"). All the others are marked as "C". The user indicated that these batches contain the term "ecology" which is too general and it is unlikely that it will contribute positively to the search results.

g. user/query 7: "optical fiber position"

The user selected only one term from a list of twenty-four terms: "optical fibres". This is one of the source terms used in the generation of the batches. The other source term, "position measurement", was not selected. The user upon further questioning indicated that this term is very general for her query, therefore she did not select it from the list.

As it will be remembered from the discussion of the results of experiment 2 in this section, owners' of the following queries similarly did not choose one of the source terms from the list of terms shown at the first step of the experiment: "computer vision detection", "hypertext internet", and "impact information technology disabled". The owner of the query "uv spectroscopy database" did not choose either of the two source terms. As it would be remembered from 8.3.1.2, in experiment 1, the owners of the queries "online information and evaluation of quality and reliability" and "polarisation mqw splitter" did not choose one or both of the two source terms.

None of the batches generated are marked as representing new ideas by the user. Two of the ten batches are marked as category B ("none of the terms in the batch are good for my search purpose"). These are Batch 1 (rank 8) and Batch 3 (rank 9). The user indicated that terms such as, "synchronisation", "time measurement" and "SONET" that appear in these batches are not useful, and these batches on the whole do not seem to be good for her search purposes.

All the other batches are marked as "C" ("some of the terms in the batch are good for my search purpose").

When quizzed after the examination of the batches about the usefulness of the source term "position measurement" that was initially not selected, the user indicated that her thoughts did not change and the term seemed too general to be useful to her search.

h. user/query 8: "expert systems object oriented development"

This user selected seven terms from a list of nineteen. These are ranked in the order of decreasing importance by the user as follows: "object-oriented methods", "knowledge representation", "expert systems", "software engineering", "logic programming", "object-oriented programming", "object-oriented databases".

The user marked one of the batches, Batch 10 (rank 6) as representing new ideas ("F"). Upon further inquiry the user stated that it is the term "multimedia systems" that he thought could be useful to him which he did not think of previously.

Batch 11 (rank 1) is marked as category A ("the batch as a whole looks good for the users search purpose"). All the other batches are marked as category D ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement"). In Batches 12 (rank 2) and 6 (rank 4) it was the term "group decision support systems", and in Batch 7 (rank 3) it was the term "deductive databases" that the user wanted to exclude from his search.

As it was observed in some of the cases of experiment 1 (8.3.1.2), although the user initially did not select one of the terms (in this case "multimedia systems") when shown in a list of unconnected terms, it was subsequently marked as a potentially useful term when shown in the context of the related terms in the batch. This seems to offer justification for presenting to the user small number of linked thesaurus terms in batches, rather than showing the terms in isolation or in a list containing several terms which are not necessarily related to each other conceptually.

8.3.2.3 The diagnostic evaluation

The first part of this exercise consists of presenting to the user five sets of 20 documents each of them containing the highest ranking 20 documents retrieved by one of the highest ranking four batches and the Okapi system (cf. 8.3.1.3). The user is then asked to go through each set of documents briefly and to make a speculation regarding which of the ten batches shown earlier might have been possibly used in generating the five sets of documents.

The purpose of this test is to find out whether the user could correlate a batch of documents with a batch of terms successfully. As noted in 8.3.1.3, this should give some indication about the usefulness of linked terms in predicting the outcome of the search process, therefore representing conceptually coherent wholes.

Clearly, as discussed in 8.3.1.3, the outcome of the retrieval process is also an effect of the search engine used in the experiment. Since the search engine used in both experiments is a best match one (cf. 7.2) and since, there are many terms that are common to all batches, it is highly likely that in many cases there could be considerable overlap of retrieved documents among the batches. Nevertheless, it is still hoped that this exercise would give some clue regarding the usefulness of the batches in helping the users in estimating what sorts of documents that are likely to be retrieved.

The owner of the query "hypertext internet" discussed in the preceding section was kind enough to agree to spend extra time to provide a detailed analysis of his results.

The user in this test expressed difficulty in correlating each set of documents with a single batch. He explained that the document sets did not seem to him as homogeneous sets retrieved by any one of the batches alone, but as if retrieved by a list of terms composed by mixing terms from a number of batches. As a consequence, the user made speculation about which batches (rather than a single batch) might have been used in retrieving the each set of documents.

For the set of documents retrieved by Batch 6 (rank 2), the user estimated that Batches 5, 4 and 6 might have been used. For the set retrieved by Batch 7 (rank 3) the user guessed that Batches 7 and 2 might have been used. For the set of documents retrieved by Batch 8 (rank 1) the user suspected that Batches 5, 4 and 10 might have been used. For the set produced by Batch 0 (rank 4), it was Batches 6, 2, 10, and 4 that were suspected. Finally for the set of documents found by Okapi, user speculated that Batches 10 and 6 might have been used.

It is clear from the above results that the user was unable to correlate the sets of documents with the correct batches used in their generation. This is not a surprising result considering that, most of the ten batches shown to the user differ from each other by just one or two terms. Since the retrieval engine used in the experiment is a partial match one, many of the top ranking documents are similar across all sets. Therefore, it was virtually impossible to tell accurately, by manual examination, the terms used in the retrieval of the document sets.

It is worth noting that, the batches used in this experiment are composed of five terms including the two source terms compared to six in the first diagnostic experiment discussed in 8.3.1.3. As a consequence of this, the overlap between terms in the batches in experiment 2 is higher compared to that of experiment 1. In other words, the batches in this experiment look more alike each other than the batches used in experiment 1. This could explain the difficulty expressed by the user in experiment 2 in identifying the sources of the retrieved sets in contrast to the relative success of the user in experiment 1.

The next step in the experiment was to ask the user to put similar batches in the same group and comment on their relationships. This is to asses the ability of the batches to represent distinct (intelligible) subject areas.

The user put Batches 2 (rank 7), 3 (rank 8) and 5 (rank 6) in the same group. Batches 8 (rank 9) and 10 (rank 10) were put in another group. Batches 6 (rank 7) and 7 (rank 8) formed another group. Batch 0 (rank 1) and 4 (rank 5) were put in another group. Batch 1 (rank 2) was left on its own without belonging to any group.

The results of this test, like the results obtained in the first diagnostic experiment discussed in 8.3.1.3, seem to support the claim that the users did indeed discern the batches as distinct wholes representing different subject areas. This is evident from the fact that the users that took part in both diagnostic experiments were able to produce groups of batches that represent similar ideas and distinguish between different groups of batches representing different ideas.

When the user participated in this exercise quizzed about the batch marked as representing new ideas (Batch 5, rank 6), the user indicated that the combination of following the terms, "database management systems", "distributed databases" and "distributed processing" made this batch interesting for him. He indicated that he was not originally conceived the search in these terms but he thought it might be interesting to try this combination. Though, he suggested that he was not in particular happy about "distributed processing" that is present in the batch which he thought might not be relevant to him.

The user stated that Batch 10 (rank 10) seemed to him the best in representing his original query.

The user explained that he marked Batches 2 and 3 as category D ("the batch contains some good terms, however there is/are term(s) in the batch that I would definitely want to exclude from my search statement") because he thought the term "concurrency control" present in these batches was not useful to him.

The user stated that Batch 4 contains the term "database theory" which he thought somewhat relevant to him, therefore it is marked as "E" ("the batch contains terms that are marginally related to my search (or of secondary importance").

He explained that he assessed Batch 7 as category G ("The batch contains term(s) that represent ideas which is/are part of the general domain of my search, however not directly useful to me"), because of the term "integrated software" which he thought was not directly relevant to his search.

The user indicated that Batch 0 is marked as "D" because of the term "relational databases", and Batch 1 because of the terms "visual databases" and "PACS", which were not relevant to him.

When quizzed about Batches 6 and 8 which are marked as "C" ("some of the terms in the batch are good for my search purpose"), the user stated that although they contain on the whole useful terms for his search, the particular combinations did not seem to describe exactly what he had in mind.

Finally, the user is asked about the usefulness of the term "computer networks". As it would be remembered from the preceding section, this is one of the source terms used in the generation of the batches that the user did not initially select it from the list of single thesaurus terms presented to him (see 8.3.2.2).

The user iterated that although "computer networks" is a related term for "internet", he thought it is not specifically designating this particular network, therefore still thought that it would be better to use specifically the term "internet" in the search.

The results of this evaluation seems to justify the use of batches in representing different shades of meaning or different sub-sections of a more general subject area. Although it seems that the users are not always consistent in the use of the eight categories in describing the batches and the reason(s) of using a particular category for different batches seems to be contradictory at times, in general it seems reasonable to conclude that, the batches seem to represent for the users that took part in the experiments distinguishable and identifiable subject areas (cf. 6.4.3).

It is probably fair to say that, although the batches on the whole repeat the same or similar terms to a large extent, the variation of one or two terms among them seems to be enough to suggest to the participants in the experiments different aspects of a more general subject area.

Similarly the documents retrieved by the batches overlap to a large degree, therefore it is not always possible to relate a given set of documents with its correct source (terms used in its retrieval), however, as argued in 8.3.1.3, this is to do as much with the partial match search engine used in the experiments, as the repetition of terms across the batches.

Overall, it can be concluded that the users seemed to find batches useful (and meaningful) in representing related areas of a general subject field, and they were able to distinguish a batch or a group of related batches from the others (despite the fact that they usually share a large number of common terms between them) in cognitive/semantic terms.

8.3.2.4 Retrieval effectiveness

The main design objective of KBS-2 that differs from KBS-1 is that, KBS-2 aims to have a high retrieval effectiveness, whereas KBS-1 aims to suggest more of areas loosely related to the user's original query. The benchmark system that is used to compare the effectiveness of KBS-2 is the Okapi system (7.2). The last stage of experiment 2 is therefore devised to compare the relevance judgements of the users for documents retrieved by KBS-2 and Okapi.

A similar methodology to the one used in experiment 1 (cf. 8.3.1.4) is employed in experiment 2. Top 20 documents retrieved by the top ranking four batches generated by KBS-2 (see appendix J for the top batches) are combined with the top 20 documents retrieved by Okapi for the same query. After duplicate documents are removed, the remaining documents are randomly ordered and showed to the owner of the original query.

The results of this experiment are given in appendix O in the tables 1 to 5. In these tables, Batches 1 to 4 (B1 to B4) are the four highest ranking batches generated by KBS-2. Batch 5 (B5) is the Okapi system. The results for these five batches are given separately for the eight users completed the experiment. Relevant (R), partially relevant (P), and not-relevant (N) documents for each batch and for each user are given at four cut-off points (5, 10, 15 and 20) as in experiment 1 (8.3.1.3).

The results presented in appendix O indicate that KBS-2's performance is comparable to Okapi in retrieval effectiveness. The best performing batches of KBS-2 have a slightly better precision in comparison to Okapi when partially relevant documents (P) are taken as relevant (R). When partially relevant documents are taken as not-relevant (N), KBS-2 performs slightly less well in comparison to Okapi. It might be suggested from this observation that, KBS-2 is better in retrieving documents that are likely to be marked as partially relevant ("P"). The reason for this could be that KBS-2 is better in finding documents that represent new ideas compared to Okapi and documents marked as "P" are those that represent new ideas to the users (this is of course a debatable point, however at least some of the "P" marked documents seem to support to this assumption).

It can be observed from the results of appendix O that, there are substantially more documents marked as relevant (R) and partially relevant (P) in experiment 2 (KBS-2) than in experiment 1 (KBS-1). This might be taken as a support for the conclusion drawn in 8.3.1.4 that WPQ is better in selecting battches that retrieve relevant (or *not* non-relevant) documents than F4. However, a slight anomaly in the ranking of the batches by WPQ should be noted. Appendix O shows that B1 performs slightly less than B2, which performs slightly less well than B4, while B3 has the worst results for the all four batches generated by KBS-2. The anomaly in this order may be a result of various heuristics (and ad hoc parameters) used in KBS-2. One should however note that, in experiment 1, F4's performance was more consistent in ordering the batches (cf. 8.3.1.4 and appendix N). It should also be noted that the number of relevance judgements obtained in the experiment is not enough to draw any statistically sound conclusions, therefore, the anomaly in the ordering of the batches could be due to the limited amount of data collected.

Similar conclusions to those drawn about KBS-1 can be drawn about KBS-2 regarding the relative positions of the relevant and partially relevant documents in the lists produced by batches 1 to 5 used in the experiment. Appendix M shows that, documents from the lower end of the ranked lists are ranked, in general, at higher positions by other batches (although there are a large number documents -- relevant and not-relevant -- ranked at similar positions by all or most of the batches). Therefore, the conclusion drawn for KBS-1 in 8.3.1.4 that it can be used to explore (conceptually) different areas of the retrieved set, can be repeated here for KBS-2.

Finally, similar to KBS-1, KBS-2 performs particularly bad in some case and as suggested in 8.3.1.4, this could be to do with the quality of the source terms matching the user's query terms in the thesaurus and the level of detail of coverage of the thesaurus in a particular subject area (see 8.4 for more on this).

8.4 Discussion of the experimental results and comparison with the CILKS project

In the following section (8.4.1), the results of the experiments 1 and 2 and general conclusions derived from these results are discussed and summarised. In the subsequent section (8.4.2), the results of the experiments performed in this project compared with the experiments performed in the CILKS project.

8.4.1 General conclusions and future research questions

General conclusions that can be drawn from the two experiments performed in this project are as follows:

- Both in terms of retrieval effectiveness and representation of subject matters small number of conceptually related terms taken from the thesaurus (i.e. a batch) seems to perform better than a larger number of unconnected terms

    - the results of experiment 1 suggests that, the individual batches perform better than a list of larger number of unconnected terms in terms of precision values

- the results of both experiment 1 and 2 suggest that users tend to find terms shown in the context of related terms in batches more effective in describing a subject area than a list of unconnected terms. Ambiguity of a thesaurus term is resolved when presented in such a context (see also 8.4.2 below)

- The retrieval effectiveness of the batches are comparable to that of the Okapi system and they tend to retrieve more documents likely to be judged "partially relevant" than Okapi

- The performance of the batches as related both to the retrieval effectiveness and ability to represent subject areas seem to be affected direcţly and strongly by the quality of the source terms. The following questions then become important in the discussion of the source terms:

  - do the source terms represent correctly the concepts present in a given query?

  - do the source terms represent all concepts that are present in a given query?

  - is the thesaurus from which the source terms are selected cover the subjects that are relevant to a given query adequately?

The results of the both experiments suggest that selection of the source terms are crucially important for the performance of the KBSs developed in this project. The source terms are used in the generation of the batches and the quality of the resulting batches and thus both the retrieval effectiveness and the ability of batches to represent subject fields are directly influenced by the two source terms selected. The retrieval effectiveness in particular is a function of the ability of the source terms in representing all concepts that are present in a given query (user input search terms). The quality of representation of a particular subject in the thesaurus on the other hand has a direct influence on the quality of the source terms.

As it can be seen in tables presented in appendix N, in experiment 1 the results for retrieval effectiveness of the user/query 6 and 7 are particularly poor. In query 6, the user did not choose either of the source terms. In query 7, the user did not choose one of the source terms. This information suggest that one or both of the source terms in these examples do not represent accurately the users' queries. This may account for the particular bad results in the above named cases. Similarly in query 7 in experiment 2, the user did not choose one of the source terms and the result for this query is the worst amongst the eight queries evaluated in this experiment. It should be noted however that, in experiment 2, not in all queries where the users did not choose one or both of the source terms, retrieval effectiveness is similarly poor. There are other factors that may influence the results and the overall retrieval effectiveness, such as the relevance of the terms linked to the source terms in a batch to the user's query and the completeness of representation of the user's original query by the batch as a whole, which may explain the differences in the retrieval effectiveness noted above.

In both of the above mentioned cases in experiment 1, the users' original queries seem to contain more than two distinct concepts. As discussed in 7.4.1, the source terms selected represent at best only two distinct concepts that may present in a query. The two source terms used in the generation of the batches basically attempts to capture only two distinct concepts that may present in user input search terms. For various reasons discussed in 7.4.1, not all user input search terms and therefore different concepts that may present in a query are always represented by the two selected source terms. Some of the user input search terms or concepts that are excluded by the source terms may appear in the batches as linked terms, but this does not need to happen always. In query 7 (experiment 1) for instance, one of the user input search terms "mqw", although present in the Inspec thesaurus as a distinct concept, was not represented in the batches. Similarly, query 4 in experiment 2, contains the terms "UV" and "database" both of which were not represented in the batches used which may account for the relatively poor performance of this query.

Lastly, the thesaurus itself may not have a specific term for a particular concept present in a query and selection of a broader term to represent that subject may result in a deteriorated performance. A likely example of such a case is the query 6 in experiment 1. In this query the user's intention was to find documents that discuss the evaluation of the *contents* of online databases in terms of quality and reliability (see 8.3.1.2). However, the Inspec thesaurus does not have a term to describe such a subject. The nearest terms that describe such a subject in the thesaurus were perhaps "information retrieval systems evaluation", "software reliability" and "software quality". Of these "information retrieval systems evaluation" and "software reliability" were selected as the source terms by the system, and "software quality" was present in the batches as a linked term. It is likely that none of these terms singly or together represent closely the concepts present in the user's original query. Some of them, e.g. "information retrieval systems evaluation", are perhaps too broad a term for this query and others take a more technical (software/hardware) point of view of evaluation.

It is arguable that the Inspec thesaurus has a bias towards technical aspects of information systems and user-oriented studies/evaluation of information systems are not represented very well. This is perhaps why in the above discussed case of query 6 in experiment 1, there was simply no term to cover the users's point of view of evaluation of the contents of online information which possibly accounts for the poor performance of the system in this case and the fact that the user did not choose either of the source terms selected by the system.

Future Research Questions

Several research questions emerge from the present study which may deserve future attention:

• Many of the heuristics used for automatic generation of the batches are database dependent, i.e. derived from examination of the structure of the Inspec thesaurus, and from the investigation of actual searches on the Inspec database (document collection).

To apply heuristics, such as, how many links need to followed before good terms appear or what are the best links for efficient navigation to a new thesaurus or document collection, the characteristics and structure of the actual thesaurus and database need to be examined. However, the experience derived from the Inspec database and thesaurus should prove to be invaluable as a model which could be extended to other thesauri and databases.

• There is a problem of processing efficiency/speed in the Oracle implementation of the Inspec thesaurus. It takes minutes rather than seconds to expand a term even for one or two levels. A faster search algorithm need to be devised in an operational interactive environment. This is partly dependent on the structure of the thesaurus (such as average number of terms associated with thesaurus terms, types of links that are in the thesaurus, as well as their purpose, characteristics etc.). However, it is on the whole a matter of effective database search techniques.

• Although, the KBSs had always two source nodes (terms), it is possible in principle to modify this to apply to the cases where there are more than two distinct concepts.

It is relatively rare that an average Okapi query involves more than two distinct aspects, corresponding to more than two distinct concepts. However, there are cases where it is necessary to represent the user's query with more than two source terms as it is observed in the experiments performed. There is evidence from the results of experiments 1 and 2 that retrieval effectiveness depends heavily on how well the query is represented by the source terms which generate the batches.

It is likely that although many queries initially involves no more than two distinct aspects, after either interaction with thesaurus or (more likely) seeing some relevant documents, this may change. Therefore, it should be useful to explore the ways to generate batches that have more than two source terms, representing three or more different aspects of a query.

• Combining descriptor terms from a thesaurus with the user's original free text query terms seems to be problematic and needs further research.

Each batch starts and ends with the same two source terms which reflect two aspects of the search subject. The two source terms are usually closely related to the original query terms, and in some case, they are identical (e.g. the query "expert systems and education" matches exactly with Inspec thesaurus terms "expert systems" and "education"). Also, the terms linked to these two source terms are such that the number of links between them and the types of the relationships involved (narrow, broader, related, lead-in etc.) are known. This structure might help us to find a better way in combining terms, essentially of different characteristics and definitely of different origin (user v. thesaurus). For example, it is plausible that number of links between a term and source terms indicates their "conceptual closeness", therefore, it can be suggested that in the search term weighting operation, such information should be taken into account (which I have not tried in my research).

• In the systems developed, the source terms were derived by matching the user input terms with the thesaurus terms. It should be worthwhile to investigate the possibility of using terms from descriptor fields of seen or relevant documents as source nodes for thesaurus navigation (generation of batches of linked terms) in a highly interactive retrieval system.

8.4.2 Comparison with the CILKS

There are certain similarities with this project and the CILKS project as already noted earlier (see especially, 6.4.2 and 7.3.1.2). The CILKS project (Jones, 1992; Jones, 1993; Jones et al., 1995) similarly used the Inspec thesaurus to expand the user input search terms. The major difference between the two projects is that, whereas in this project the query expansion is done automatically by the system, in CILKS it is done partly by the user. The user is presented with a list of terms matching the user's original search terms automatically by the system in CILKS (7.3.1.2). The user can then follow the equivalence, hierarchical, and associative relationships that a term may have in the Inspec thesaurus manually and choose terms from the initial matching thesaurus terms or the terms linked to them with one of the available relationships to include in the search statement.

It should be worthwhile to compare the systems implemented in this project with the system designed as part of the CILKS project, in terms of users' term selection and use behaviour, as these two projects represent two distinct and contrasting approaches to IRS design practice. This should give some idea about the usefulness of the semiotic principles which guided the design and evaluation processes of the present study. The purpose of this section is, therefore, to compare the results of the evaluation experiments reported in section 8.3 with some observations from the CILKS project regarding the user search behaviour.

I was involved with one of the evaluation experiments of the CILKS project. I had organised the sessions with the individual users that took part in the experiment, conducted the experiment and collected data. This gave me the opportunity to observe the users' interaction with the thesaurus and their navigation and term selection behaviour. This first hand experience was useful in my own project in shaping my thinking regarding the ways of using thesaurus as a search aid. In what follows below, I will attempt to illustrate some of the problems experienced by the participants in the above mentioned CILKS experiment and how they are overcome by the KBSs developed as part of my Ph.D. project.

One of the main problems with the CILKS system is that, if the initial set of matched terms are not good, users find it difficult to employ appropriate search tactics to find their way in the thesaurus. They do not seem to be able to use the links that eventually lead to better terms. For example, an experienced searcher might choose a broader term to go up the hierarchy and then follow the related and narrower links of the thesaurus to find better terms. An inexperienced user on the other hand is more likely not to be able to follow this sort of tactical links and might therefore not be able to explore the knowledge space fully.

To illustrate this sort of user difficulty with thesaurus navigation, consider the following example from the CILKS experiment: one of the queries was "cd-rom standards" and although the initial matched thesaurus terms include the term "standards", the user did not choose it, stating that it is too general, although he did not find any better term in the matched list which covers this aspect of his query.

The KBS devised in the present project would however put this term in context by displaying to the user all the terms that link "standards" to "cd-roms" and enable the user to make an informed decision on which terms to choose for query expansion/formulation. The KBS developed would inform the user on the overall structure of the domain as defined by his query by displaying the various aspects of it.

The following list of batches illustrates, how the domain defined by a statement such as "cd-rom standards" is divided into sub-classes automatically by the KBS. It also shows how the system puts the terms in context by generating sets of conceptually related terms. This removes any ambiguity that a term might have when displayed on its own to the user.

Batch 0

0 standards
1 software portability
2 software engineering
3 computer software
4 firmware
5 read-only storage
6 CD-ROMs

Batch 1

0 standards
1 standardisation
2 code standards
3 data handling
4 document image processing
5 electronic publishing
6 CD-ROMs

Batch 4

0 standards
1 telecommunication standards
2 telecommunication
3 telecommunication services
4 multimedia systems
5 optical publishing
6 CD-ROMs

Batch 5

0 standards
1 telecommunication standards
2 television standards
3 television
4 video recording
5 video and audio discs
6 CD-ROMs

Batch 31

0 standards
1 code standards
2 data handling
3 text editing
4 information science
5 information retrieval systems
6 CD-ROMs

Batch 32

0 standards
1 code standards
2 data handling
3 text editing
4 information science
5 information storage
6 CD-ROMs

Batch 75

0 standards
1 electronic data interchange
2 data communication systems
3 ISDN
4 multimedia systems
5 electronic publishing
6 CD-ROMs

Batch 85

0 standards
1 electronic data interchange
2 data handling
3 document image processing
4 records management
5 information retrieval
6 CD-ROMs

Batch 106

0 standards
1 open systems
2 computer architecture
3 computers
4 digital storage
5 storage media
6 CD-ROMs

Batch 107

0 standards
1 open systems
2 computer architecture
3 memory architecture
4 digital storage
5 read-only storage
6 CD-ROMs

Batch 115

0 standards
1 protocols
2 technical office protocol
3 office automation
4 records management
5 information retrieval
6 CD-ROMs

Batch 116

0 standards
1 protocols
2 technical office protocol
3 office automation
4 records management
5 information storage
6 CD-ROMs

Batch 120	Batch 121
0 standards	0 standards
1 protocols	1 protocols
2 computer interfaces	2 computer interfaces
3 computer comm. software	3 computer peripheral equipment
4 online front-ends	4 digital storage
5 information retrieval systems	5 read-only storage
6 CD-ROMs	6 CD-ROMs

By using the technique described in 7.3 to calculate the value of relatedness between the query as a whole and the terms in the batches, the KBS generate candidate batches for query expansion/formulation. This method proved to be useful for both finding new relevant documents and as an aid for conceptually describing the thesaurus space (i.e. structured navigation).

The above example of "cd-rom standards" illustrates that, many queries in interactive environments are stated very vaguely. The batches of terms generated by the KBS however remove this vagueness by dividing the search subject into smaller categories which approximate different aspects of a subject field.

Another important problem encountered in the CILKS system is that, the users seem to find it difficult to judge the contents of terms. In a large knowledge representation structure, such as a thesaurus, the domain covered usually is very wide. As a result of this, contents of terms are highly context sensitive (cf. 7.1.3), and same or similar terminology may be used in a totally different sense in different domains. To illustrate this from the CILKS experiment consider one of the queries: "user interface evaluation". In this query the user chose "function evaluation" from the initial matching terms, most probably thinking that it has something to do with the functional evaluation of user interfaces. Actually, the term is from "Mathematics" and completely irrelevant to the user's search topic. This is not the only example of this type of mistake. In fact, another user whose query was "evaluation of information systems" made the same mistake and chose "function evaluation". Note that as indicated in the preceding section, the Inspec thesaurus does not have a specific term for evaluation of information systems/interfaces from a user-centred perspective and this defect of the thesaurus causes the users searching in this area to become frustrated or select wrong terms. There are more examples of this kind and the quality of the terms chosen by the users in the CILKS experiment, in general, is dubious.

One way to overcome this problem is to use semantic net type structure to display the terms in context as noted above. As discussed in 8.4.1 and elsewhere in this chapter, when users are presented with a set of terms conceptually related forming a coherent set, their judgements of the usefulness of the terms in the set seem to be more consistent. Also, they seem to assess value of a set of terms as a whole for searching more easily and accurately. The findings of experiment 1 and 2 also indicate that, system generated sets of linked terms (batches) differentiate from each other conceptually, i.e. users are able to assign different contents to different sets (cf. 8.4.1, 8.3). This is evident from the ability of the users who took part in the experiments to assign different categories to different sets when asked to assess the value of the batches in relation to their query using the eight categories provided.

The number of descriptor terms used for searching seems to be have an effect on the quality of the results significantly. The results of experiment 1 (8.3.1.4) suggests that, precision deteriorates with the increase in the number of descriptor terms used in searching. This is really not a surprise, considering that at least for the top portion of the list (which is possibly the most important portion in an interactive environment), the level of noise is greater when a lot of descriptors terms are used. Evidence from CILKS is that users tend to select large numbers of thesaurus terms, representing different aspects of their query. For example, a user whose query statement was "sensor chemical gas electro- optical" has selected some 80 terms which include; artificial intelligence, rivers, lakes, physical chemistry, in addition to what one expects from this sort of query, such as; gas sensors, electrochemical analysis, sensor fusion, spectrochemical analysis, etc.

The above might be an extreme case, however, there are several examples of this sort of search behaviour in CILKS. The evidence from the CILKS experiment is that many searches are quite broad and thesaurus navigation usually helps one to define it better. Yet this need to be done not by lumping every aspect of the search subject in one big query, but rather arranging them into suitable sets of conceptually related terms (i.e. batches). This would eliminate the noise which is caused by mixing different aspects of a subject in a single big query statement by limiting the number of terms that can be grouped together and arranging them into some logically and conceptually coherent clusters (batches).

It is likely that, it would be even more beneficial if the user can interact with the batch generation process, e.g. choose source terms, able to select/deselect the terms that are linked (in the same batch) for inclusion/exclusion from the search, etc.

In a highly interactive environment, it should be desirable to have a facility which enables the user to find out the links between any given two terms. It is plausible to suggest that one of the ways a human searcher make use of a thesaurus is to look for the links between search terms. Obviously, when one gets this information (i.e. what terms link two given terms), it substantially increases the searcher's understanding of the concepts involved in the search statement and puts the search/thesaurus terms in context and removes any ambiguity that they may have.

Chapter 9
Conclusions and Future Research

This research had closely related dual objectives. It was felt that, before retrieval systems design and evaluation tasks could be attempted, it was first necessary to clarify many of the fundamental concepts and assumptions in use in IR theory. The reason for this is the belief that, any design process should rest, as much as possible, on concepts and presumptions *explicitly* defined and acknowledged. The first objective of this research was, therefore, to analyze the document retrieval process and explicate the basic IR concepts from the point of view of semiotics which was chosen, for the reasons explained in the introductory chapter, as the main methodological framework in this project. The ultimate objective was, of course, to develop retrieval systems that conform with the theoretical analysis of the retrieval situation.

It should be noted however that, the relationship between the theory and practice was not uni-directional, but bi-directional. In other words, while the theory guided the design and evaluation processes, the practical experience gained from the systems design process in the course of this project had been fed into the theory development. Therefore, each task proved to be indispensable for sound development of the other.

The most important result of the semiotic model developed was the explication of the distinction between the knowledge production and transfer functions of document retrieval. The consequence of this finding was the conceptualization of the retrieval process as a dynamic and complex interplay between knowledge production and transfer tasks. This was translated into actual design practice as the system objective of dividing a general search area defined by a user query into smaller more specific ones. In practice this was achieved by generating clusters of terms linked in the Inspec thesaurus. Each cluster or batch of terms was conceived as representing a part of the general search area defined by the user. The purpose here was to enable the user to identify new search areas from the term information contained in the batches.

9.1 Main results of the Evaluation Experiments

The evaluation experiments performed aimed at finding out whether the batches were actually effective in defining search areas related to the original user queries and whether they were useful in pointing new areas which were potentially relevant to the users. The following are the main results of the experiments:

- the batches were useful in representing search domains relevant to the users' queries

- in many cases the batches represented new ideas or new search domains to the users

- the batches were useful in removing ambiguities associated with single thesaurus terms

Also a number of questions related to the retrieval effectiveness of the knowledge based systems designed were asked in the experiments. The results indicated that:

- the knowledge based systems had similar effectiveness in terms of precision as the Okapi system

- the knowledge based systems tended to find more partially relevant documents than the Okapi system

More detailed analysis of the results are given in 8.3, and a summary of the results can be found in 8.4.1. and 8.4.2. However, as a general conclusion it can said that, the users participated in the experiments seemed to find the batches useful in explicating their queries. In many cases the users seemed to be interested in following new ideas indicated by the batches or the terms in the batches. This last point suggests that, many of the users had interests in more than one area of the general search space indicated by their initial queries. If we were to generalize the results further, it might be possible to make the following statement about end-user searching in IR: users of IR systems could be expected in many cases to have more than one distinct but interrelated queries, and it would be a good service if retrieval systems help the users to explicate the different aspects of their query.

It can be suggested from the findings of this project that, it would be even more beneficial if the users were involved in the source terms selection and batch generation processes. The knowledge based systems were really designed with the interactive end-user searching in mind. It seems to me that the ability of finding the links between any given two terms automatically is a very effective way of exploring new ideas and relationships in a thesaurus. It is desirable for the searcher in an interactive environment to be able to interfere with the source terms selection and batch generation processes and select/deselect the system suggested terms and links.

9.2 Semiotics and Future Research Directions

Semiotic analysis of document retrieval presented in chapters 4, 5, and 6 provided a rich framework for understanding the retrieval situation. Obviously, only a few of the ideas that emerged from the semiotic analysis could be tested within the limits of the present project. Many of the ideas are therefore left for future research to be tested. Some of the ideas emerged from the semiotic view of the IR interaction which may deserve future research attention will be discussed briefly below.

As noted earlier, the most important theoretical distinction that the semiotic view of IR introduces is that of between knowledge production and transfer functions of document retrieval systems. The users that took part in the experiments performed as part of the project indicated in many cases that, batches as a whole or individual terms constituting the batches suggested new ideas to them. However, it was not tested in this project whether, those batches marked as representing new ideas by the users embody genuinely new relationships between thesaurus terms (which stand for important concepts in specific knowledge domains), or they just represent relationships already well established in the public domain, however, new to the particular user.

If it could be established by future research that, new (publicly unknown) ideas could be produced by following term relationships in a thesaurus, this technique could be deployed in cases where the creation of new knowledge has a priority over transfer of established knowledge to assist the users in their creative activities. The next step would then be to devise experiments to find out the most fruitful ways of integrating the two fundamental language games in IR (i.e. knowledge production and transfer) in interactive end-user search systems. It is likely that, different user groups involve with knowledge production and transfer tasks to varying degrees.

It would therefore be useful to know more about how knowledge is produced and used by different knowledge communities to design information retrieval systems that can serve their specific needs.

The semiotic view presented in this dissertation portrays IR interaction as a dynamic and undetermined process, where, users' information needs change and branch out all the time. The case of individual user with a static and determined need that can be anticipated is perhaps not the norm in IR, but a special case. This situation corresponds to the case of didactics or knowledge transfer in the semiotic model developed. In the cases of knowledge production however, information relevant to a user dynamically changes. This is perhaps a more accurate description of real retrieval processes where each document (or other types of information items) seen causes the users' interests to diverge and fork out to new directions.

It is perhaps impossible in such cases to anticipate the information needs of the users. Although, in a limited number of cases it might be possible to foresee future uses or potentials of documents and devise indexing and classification schemes accordingly, for most of the time such devices have to rely on what is already known in a subject domain to record relationships between documents. The semiotic analysis of the retrieval situation suggests that, knowledge production takes place in a social and cultural realm, therefore, potential uses of documents are not determined by a single individual. It is rather determined collectively. For this reason, even if indexing schemes are developed with utmost care and documents are indexed diligently, they may still lag behind the developments in knowledge domains that they aim to serve. Therefore, use of a relatively static knowledge structure, such as a thesaurus may have a rather restricted use for the purpose of knowledge production.

It may be possible, however, to device other mechanisms to help the users in their knowledge production activities. Since knowledge production is a collective, social process which involves interaction and communication among the participants, it may be useful to integrate search and communication functions in one application. A retrieval system which combines both functions would enable a group of people with shared research interests to participate in knowledge production more effectively by enabling them to exchange ideas, share newly discovered documents, and reuse past searches of the group members. Such a system would be useful in helping the users in conceiving, discussing, and negotiating new ideas and statements, which could lead to production of new knowledge.

References

Aitchison, J. (1972). *General linguistics.* London: Teach Yourself Books.

Andersen, P.B. (1986). Semiotics and informatics: computers as media. In: P. Ingwersen, L. Kajberg, A.M. Pejtersen (eds.), *Information Technology and Information Use* (pp. 64-97). Taylor Graham.

Andersen, P.B. (1990). *A theory of computer semiotics -semiotic approaches to construction and assessment of computer systems.* Cambridge: Cambridge University Press.

Anderson, J.R., & Bower, G.H. (1973). *Human Associative memory.* Washington DC:Winston.

Ashford, J., & Willet, P. (1988). *Text retrieval and document databases.* Bromley: Chartwell-Bratt.

Austin, J.L. (1962). *How to do things with words.* Oxford: Oxford University Press.

Bangura, A.K (1994). Semantic representation with and without logic. *Languages of design, 2,* 153-161.

Bar-Hillel, Y. (1964). *Language and Information -selected essays on their theory and application.* Reading, Mass.: Addison-Wesley.

Barthes, R. (1967). *Elements of Semiology.* Hill and Wang.

Barthes, R. (1972). *Mythologies.* London: Cape.

Barthes, R. (1985). *The fashion system.* London: Cape.

Bates, M.J. (1979). Information search tactics. *Journal of the American Society for Information Science, 30,* 204-214.

Bawden, D (1990). *User oriented evaluation of information systems and services.* Hants: Gower Publishing Co. Ltd.

Bawden, D. (1986). Information systems and the stimulation of creativity. *Journal of information Science, 12,* 203-216.

Beardon, C. (1994). Computers postmodernism and the culture of the artificial. *Artificial Intelligence and Society, 8,* 1-16.

Bechtel, W. (1988). *Philosophy of Mind -an overview for cognitive science.* Hillsdale, N.J.: L. Erlbaum Associates.

Belkin, N.J. (1980). Anomalous states of knowledge as a basis for information retrieval. *The Canadian Journal of Information Science, 5,* 133-143.

Belkin, N.J., & Robertson, S.E. (1976). Information science and the phenomenon of information. *Journal of the American Society for Information Science, July-August,* 197-204.

Belkin, N.J., Oddy, R.N., & Brooks, H.M. (1982a). ASK for information retrieval: Part I. background and theory. *Journal of Documentation, 38,* 61-71.

Belkin, N.J., Oddy, R.N., & Brooks, H.M. (1982b). ASK for information retrieval: Part II. results of a design study. *Journal of Documentation, 38,* 145-164.

Berry, M. (1977). *An introduction to systemic linguistics, vol. 1: Structures and Systems.* Batsford.

Blair, D.C. (1992). Information retrieval and the philosophy of language. *Computer Journal, 35,* 200-207.

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

Bloomfield, L. (1933). *Language.* Holt, Rinehart and Winston.

Bookstein, A. (1989). Set oriented retrieval. *Information Processing & Retrieval, 25,* 465-475.

Bookstein, A., & Kraft, D. (1977). Operations research applied to document indexing and retrieval decisions. *Journal of the ACM, 24,* 418-427.

Bookstein, A., & Swanson, D. (1974). Probabilistic models for automatic indexing. *Journal of the American Society for Information Science, 25,* 312-318.

Bookstein, A., & Swanson, D. (1975). A decision theoretic foundation for indexing. *Journal of the American Society for Information Science, 26,* 45-50.

Borko, H. (1964). Measuring the reliability of subject classification by men and machines. *American Documentation,* 268-273.

Braekevelt, P. & Wade, S. (1995). Use of IR_framework in the teaching of information retrieval. *The New Review of Document & Text Management, 1,* 237-251.

Brajnik, G., Guida, G., & Tasso, C., (1986). An expert interface for effective man-machine interaction. In: L. Bolc & M. Jarke (Eds.), *Cooperative interfaces to information system* (pp. 259-308). Berlin: Springer-Verlag.

Brier, S. (1992). A philosophy of science perspective - on the idea of a unifying information science. In: P. Vakkari, & B. Cronin (Ed.) (1992). *Conceptions of Library and Information Science - Historical, Empirical and Theoretical Perspectives: Proceedings of the International Conference Held for the Celebration of the 20th Anniversary of the Department of Information Studies, University of Tampera, Finland, 26-28 August 1991* (pp. 97-108). London: Taylor Graham.

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

Brooks, H.M, Daniels, P.J., & Belkin, N.J. (1986). Research on information interaction and intelligent provision mechanisms. *Journal of Information Science, 12,* 37-44.

Brown, C.H. (1974). *Wittgensteinian Linguistics.* Mouton.

Buckley, C. Salton, G., Allan, J., Singhal, A. (1995). Automatic query expansion using SMART. In: *Proceedings of the Third Text Retrieval Conference (TREC-3).* NIST Special Publication 500-225.

Chao, Y.R. (1968). *Language and symbolic systems.* Cambridge UP.

Chen, H., (1992). Knowledge-based document retrieval: framework and design. *Journal of Information Science, 18*, 293-314.

Chen, H., & Lynch, K.J. (1992). Automatic construction of network of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics, 22*, 885-902.

Chen, H., & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval systems design. *Information Processing & Management, 27*, 405-432.

Chen, H., Lynch, K.J., Bashu, K., & Ng, T.D. (1993, April). Generating, integrating and activating thesauri for concept-based document retrieval. *IEEE Expert*, 25-34.

Chen, H., Ng, T.D., Martinez J., Schatz, B.R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. *Journal of the American Society for Information Science, 48*, 17-31.

Chen, H., Yim, T., Fye, D., Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science, 46*, 175-193.

Cherry, C. (1968). *On human communication.* Cambridge, Mass.: The M.I.T. Press.

Chiaramella, Y., & Defude, B. (1987). A prototype of an intelligent system for information retrieval:IOTA. *Information Processing & Management, 23*, 285-303.

Chong, A. (1989). Topic: a concept-based document retrieval system. *Library Software Review, 8*, 281-284.

Cleverdon, C. (1967). The cranfield tests on index language devices. *Aslib proceedings, 19*, 608-620.

Cohen, P.R., & Feigenbaum, E.A. (1982). *The handbook of artificial intelligence. Volume 3.* CA: William Kaufmann.

Cohen, P.R., & Kjeldsen, R. (1987). *Information Processing & Management, 23*, 255-268.

Cooper, W.S. (1973). On selecting a measure of retrieval effectiveness. Part 1. The subjective philosophy of evaluation. *Journal of the American Society for Information Science, 24*, 87-100.

Crane, D. (1972). *Invisible colleges.* University of California press.

Crystal, D. (1971). *Linguistics.* Harmondsworth: Penguin.

Croft, B.W., & Harper, D.J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation, 34*, 285-295.

Croft, B.W. & Thompson, R.H. (1987). I^{3}R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science, 38*, 389-404.

Cronin, B., & Hert, C.A. (1996). Scholarly foraging and network discovery tools. *Journal of Documentation, 51*, 388-403.

Crouch, C.J. (1990). An approach to the automatic construction of global thesauri. *Information Processing & Management, 26*, 629-640.

Crouch, C.J., & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. In: *Proceedings of the Fifteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24, 1992*, (pp. 77-88).

Davies, R. (1989). The creation of new knowledge by information retrieval and classification. *Journal of Documentation, 45*, 273-301.

Deely, J. (1985). *Semiotics 1984.* University press of America.

Deleuze, G., & Guattari, F. (1987). *A thousand plateaus: capitalism and schizophrenia.* Minneapolis: University of Minnesota Press.

Deleuze, G, and Guattari, F. (1977). *Anti-Oedipus - capitalism and schizophrenia.* New York: Viking Press.

Deogun, J.S., & Raghavan, V.V. (1986). User-oriented document clustering: a framework for learning in information retrieval. In: *ACM Conference on Research and Development in Information Retrieval, Pisa, Italy* (pp. 157-163).

Derrida, J. (1974). *Of Grammatology.* Baltimore: The John Hopkins University Press.

Dervin, B. (1977). Useful theory for librarianship - communication, not information. *Drexel Library Quarterly, 13*, 16-32.

Dervin, B., & Nilan, M. (1986). Information needs and uses. In: M. Williams (Ed.), *Annual Review of Information Science and technology*, (vol. 21, pp. 3-33). White Plains, NY:Knowledge Industry.

Desouza, C.S. (1993). The semiotic engineering of user-interface languages. *International Journal of Man Machine Studies, 39*, 753-773.

Doede, N. (1972). *The Meaning of Information.* The Hague: Mouton & Co.

Doszkocs, T.E. (1978). AID - an associative interactive dictionary for an on-line searching. *Online review*, 2, 163-173.

Doszkocs, T.E. (1983). CITE NLM: Natural language searching in an online catalog. *Information Technology and Libraries, 2*, 364-380.

Doszkocs, T. E. (1986). Natural Language processing in information retrieval. *Journal of the American Society for Information Science, 37*, 191-196.

Doszkocs, T.E., & Rapp, B.A. (1979). Searching MEDLINE in English: a prototype user interface with natural language query, ranked output and relevance feedback. In: *Information Choices and Policies, Proceedings of the 42nd ASIS Annual Meeting, Minneapolis, Minnesota, Oct. 14-18*, (pp. 132-139). NY: Knowledge Industries Publications Inc.

Dreyfus, H.L. (1992). *What computers still can't do -a critique of artificial reason.* London: MIT Press.

Eco, U. (1976). *A Theory of semiotics.* Bloomington: Indiana University Press.

Eco, U. (1979). *The role of the reader-explorations in the semiotics of text.* Bloomington: Indiana University Press.

Eco, U. (1984). *Semiotics and the philosophy of language.* London: Macmillian.

Efthimiadis, E.N. (1990). Online searching aids: a review of front-ends, gateways and other interfaces. *Journal of documentation, 46*, 218-262.

Efthimiadis E.N. (1992). *Interactive query expansion and relevance feedback for document retrieval systems.* Unpublished Ph.D dissertation, City University, London.

Ellis. D. (1992). The physical and cognitive paradigms in information retrieval research. *Journal of Documentation, 48*, 45-64.

Erman, L.D., Hayes-Roth, F., Lesser, V.R., & Reddy, D.R. (1980). The Hearsay II speech understanding system: integrating knowledge to resolve uncertainty. *ACM Computing Surveys, 12*, 213-253.

Everitt, B. (1980). *Cluster analysis* (2nd ed.) London: Heinemann.

Feyerabend, P. (1978). *Against Method.* London: Verso.

Fidel, R. (1985). Moves in online searching. *Online Review, 9*, 61-74.

Fidel, R. (1986). Towards expert systems fro the selection of search keys. *Journal of the American Society for Information Science, 37*, 37-44.

Findler, N. (Ed.) (1979). *Associative networks: the representation and use of knowledge by computers.* New york: Academic Press.

Fiske, J. (1982). *Introduction to communication studies.* London: Methuen.

Fitzpatrick, L., & Dent, M. (1997). Automatic feedback using past queries: social searching? In: *Proceedings of the Twentieth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, Pennsylvania, July 27-31, 1997* (pp. 306-313).

Froehlich J.T. (1989a). The foundations of Information Science in social epistemology. In: *Proceedings of the Twenty Second Annual Hawaii International Conference on System Sciences* (pp. 306-315). Washington DC: IEEE Computer Science Press.

Froehlich J.T. (1989b). Relevance and the relevance of social epistemology. In: *Information Knowledge Evolution: Proceedings of the 44th FID Conference, Helsinki* (pp. 55-63).

Frohmann, B. (1992). The cognitive viewpoint in IR. *Journal of Documentation, 48,* 365-386.

Frohmann, B. (1990). Rules of indexing: a critique of mentalism in information retrieval theory. *Journal of Documentation, 46,* 81-101.

Frohmann, B. (1992). The Power of images - a discourse analysis of the cognitive viewpoint. *Journal of Documentation, 48,* 365-386.

Frohmann, B. (1994). Communication technologies and the politics of postmodern information science. *Canadian Journal of Information and Library Science, 19,* 1-22.

Fuhr, N. & Buckley, C. (1989). Probabilistic indexing from relevance feedback data. In: *Proceedings of the Thirteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Brussels, Belgium, Sept. 5-7, 1989* (pp. 45-62).

Furnas, G.W., Landauer, T.K., Gomez, L.M., & Dumais, S.T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30,* 964-971.

Garvey, W. D. (1979). *Communication - the essence of science: facilitating information exchange among librarians, scientists, engineers and students.* Oxford: Pergamon.

Gibbons, R. (1992). *A primer in game theory.* Hemel Hempstead: Harvester.

Gleason, H.A. (1961). *An introduction to descriptive linguistics.* Holt, Rinehart and Winston.

Godel, K. (1962). *On formally undecidable propositions of 'principia mathematica' and related systems.* Oliver & Boyd.

Goffman, W. (1969). An indirect method of information retrieval. *Information Storage and Retrieval, 4,* 361-373.

Goker, A.S. (1994). *An investigation into the application of machine learning in information retrieval.* Unpublished Ph.D dissertation, City University, London.

Grice, H.P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Guntzer, U., Juttner, G., Seegmuller, G. & Sarre, F. (1989). Automatic thesaurus construction by machine learning from retrieval sessions. *Information Processing & Management, 25,* 265-273.

Halloran, J. D. (1983). Information and communication: information is the answer, but what is the question? *Journal of Information Science, 7,* 159-167.

Hammarstrom, G. (1975). *Linguistic units and items.* Berlin: Springer-Verlag.

Hancock-Beaulieu, M. (1990). Evaluating the impact of an online library catalogue on subject searching behaviour at the catalogue and at the shelves. *Journal of Documentation, 46,* 318-338.

Hancock-Beaulieu, M., Robertson, S.E., & Neilson, C. (1991). Evaluation of online catalogues - eliciting information from the users. *Information Processing & Management, 27*, 523-532.

Harper, D.J. & van Rijsbergen, C.J. (1978). An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation, 34*, 189-216.

Harter, S.P. (1975a). A probabilistic approach to automatic keyword indexing, Part I: on the distribution of speciality words in technical literature. *Journal of the American Society for Information Science, 26*, 197-206.

Harter, S.P. (1975b). A probabilistic approach to automatic keyword indexing, Part II: an algorithm for probabilistic indexing. *Journal of the American Society for Information Science, 26*, 280-289.

Harter, S.P., & Cheng, Y.R. (1996). Colinked descriptors: improving vocabulary selection for end-user searching. *Journal of the American Society for Information Science, 47*, 311-325.

Harter, S.P. & Peters, A.R. (1985). Heuristics for online information retrieval: a typology and preliminary listing. *Online Review, 9*, 407-424.

Hervey, S. (1982). *Semiotic perspectives.* George Allen & Unwin.

Hjørland, B. (1992). The concept of 'subject' in information science. *Journal of Documentation, 48*, 172-200.

Hjørland, B. & Albrechtsen, H. (1995). Toward a new horizon in information science: domain analysis. *Journal of the American Society for Information Science, 46*, 400-425.

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

Hjørland, B. (1998). Information retrieval, text composition and semantics. In: *Proceedings of the 1998 ISKO conference* (in press).

Hockett, C.F. (1969). *A course in modern linguistics.* Macmillian.

Hoel, I.A.L (1992). Information science and hermeneutics - should information science be interpreted as a historical and humanistic science? In: P. Vakkari, & B. Cronin (Ed.) (1992). *Conceptions of Library and Information Science - Historical, Empirical and Theoretical Perspectives: Proceedings of the International Conference Held for the Celebration of the 20th Anniversary of the Department of Information Studies, University of Tampera, Finland, 26-28 August 1991* (pp. 69-81). London: Taylor Graham.

Ingwersen, P.O. & Pors, N.O. (Eds.) (1996). *Proceedings of CoLIS 2, Second International Conference on Conceptions of Library and Information Science: Integration in Perspective, Oct. 13-16, 1996.* Copenhagen: The Royal School of Librarianship.

Jameson, F. (1972). *The Prison house of language -a critical account of structuralism and Russian formalism.* Princeton University Press.

Jones, S. (1992). *Cilks progress report.* Unpublished internal report, City University, London.

Jones, S. (1993). A thesaurus data model for an intelligent system. *Journal of Information Science, 190,* 167-178.

Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J. & Walker, S. (1995). Interactive thesaurus navigation: Intelligence rules OK? *Journal of the American Society for Information Science, 46,* 52-59.

Karamuftuoglu, M. (1996). Semiotics of documentary information retrieval systems. In: *Proceedings of CoLIS 2, Second International Conference on Conceptions of Library and Information Science: Integration in Perspective, Oct. 13-16, 1996* (pp. 85-97). Copenhagen: The Royal School of Librarianship.

Karamuftuoglu, M. (1997). Designing language games in Okapi. *Journal of Documentation, 53,* 69-73.

Karamuftuoglu, M. (in press). Collaborative information retrieval: Toward a social informatics view of IR interaction. *Journal of the American Society for Information Science.*

Karlgren, J. (1993). Sublanguages and registers: a note on terminology. *Interacting with Computers, 5,* 348-350.

Kim, Y.W., & Kim, J.H. (1990). A model of knowledge based information retrieval with hierarchical concept graph. *Journal of Documentation, 46,* 113-136.

Kristeva, J. (1989). *Language - The unknown: an initiation into linguistics.* London: Harvester Wheatsheaf.

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

Lancaster, F.W. (1979). *Information retrieval systems -characteristics, testing and evaluation (2nd ed.).* New York: Wiley.

Lancaster, F.W. (1986). *Vocabulary control for information retrieval (2nd ed.).* Information Resources Press.

Lesk, M.E. (1969). Word-word associations in document retrieval systems. *American Documentation, 20,* 27-38.

Lewis, D.D. (1991). Natural language processing and text classification: position paper. In: *Proceedings of the Workshop on Future Directions in Text Analysis, Retrieval, and understanding, Oct. 10-11, 1991, Chicago, IL* (pp. 52-57).

Lewis, D.D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In: *Proceedings of the Fifteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 1992* (pp. 37-50).

Lewis, D.D. (1995). Evaluating and optimizing autonomous text classification systems. In: *Proceedings of the Eighteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, July 9-13, 1995* (pp. 246-254).

Lewis, D.D., Croft, W.B., & Bhandaru, N. (1989). Language-oriented information retrieval. *International Journal of Intelligent Systems, 4*, 285-318.

Liebenau, J., & Blackhouse, J. (1990). *Understanding information - an introduction.* London: Macmillian.

Lyons, J. (1968). *Introduction to theoretical linguistics.* Cambridge: Cambridge University Press.

Lyons, J. (1977). *Semantics - volume 1.* Cambridge: Cambridge University Press.

Lyotard, J. -F. (1984). *The postmodern condition - a report on knowledge.* Manchester: Manchester University Press.

Machlup, F. (Ed.) (1983). *The study of information -interdisciplinary messages.* New York: Wiley.

Malmberg, B. (1976). *Structural linguistics and human communication.* Berlin: Springer-Verlag.

Marcus, R.S. (1983). Computer-assisted search planning and evaluation. In: *Proceedings of the 46th ASIS annual meeting, vol. 20, October 1983* (19-21).

Maron, M.E., & Kuhns, J.L. (1960). On relevance, probabilistic indexing, and information retrieval. *Journal of the ACM, 7*, 216-244.

Maltz, D. & Ehrlich, K. (1995). Pointing the way: active collaborative filtering. *Human Factors in Computing Systems (CHI '95) - Proceedings, vol. 1* (pp. 202-209). New York: ACM.

Miller, D. W. (Ed.) (1983). *A pocket Popper.* London: Fontana.

Miller, G.A., Beckwith, R., Felbaum, D., Gross, D., & Miller, K. (1990). *Introduction to Wordnet - an online lexical database.* Princeton University, Cognitive Science Laboratory, Technical report No. 43.

Minsky, M. (ed.) (1968). *Semantic information processing.* Cambridge, MA: MIT Press.

Moi, T. (Ed.) (1986). *The Kristeva reader.* Oxford: Basil Blackwell.

Moles, A. (1966). *Information theory and esthetic perception.* Urbana, Chicago: University of Illinois press.

Nake, F. (1994). Human computer interaction -signs and signals interfacing. *Languages of design, 2*, 193-205.

Nii, H.P. (1986). Blackboard systems: the blackboard model of problem solving and evolution of blackboard architectures. *AI Magazine, 7*, 38-53.

Norris, C (1982). *Deconstruction theory & practice.* London: Routledge.

O'Connor, B.C. (1993). Browsing -a framework for seeking functional information. *Knowledge: Creation, Diffusion, Utilization, 15*, 211-232.

Paice, C.D. (1991). A thesaural model of information retrieval. *Information Processing & Management, 27,* 433-447.

Parsons, T. (1967). *The social system.* Glencoe, Il.: Free Press.

Peat, H.J., & Willet, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42,* 378-383.

Politt, A.S. (1987). CANSEARCH: An expert systems approach to document retrieval. *Information Processing & Management, 23,* 119-138.

Politt, A.S. (1988). A common query interface using MenUSE - a menu based user interface search engine. In: *Proceedings of the 12th International Online Meeting.* Dec. 8-10, 1988, London (pp. 445-457). Oxford: Learned Information.

Posner, R. (1992). Origins and development of contemporary syntactics. *Languages of Design, 1,* 37-50.

Poster, M. (1984). *Foucault, Marxism and History -mode of production versus mode of information.* London and New York: Polity Press.

Quillian, M.R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing* (pp. 216-270). Cambridge, MA: MIT Press.

Rada, R., & Bicknell, E. (1989). Ranking documents wit a thesaurus. *Journal of the American Society for Information Science, 40,* 304-310.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *III Transactions on Systems, Man and Cybernetics, 19,* 17-30.

Raghavan, V.V. & Jung, G.S. (1989). A machine learning approach to automatic pseudo-thesaurus construction. In: *4th International Symposium on Methodologies for Intelligent Systems, Charlotte, North Caroline, Oct. 12, 1989.* New York: Elsevier Science Publishing.

Rapoport, A. (1960). *Flights, games, debates.* University of Michigan, Center for Research on Conflict Resolution.

Rapoport, A. (1966). *Two person game theory - the essential ideas.* The University of Michigan Press.

Rasmusen, E. (1994). *Games and information - an introduction to game theory.* Oxford: Blackwell.

Robertson, S.E. (1977a). The probability ranking principle in IR. *Journal of Documentation, v 33,* 294-304.

Robertson, S.E. (1977b). The probabilistic character of relevance. *Information Processing & Management, 13,* 247-251.

Robertson, S.E. (1979). Indexing theory and retrieval effectiveness. *Drexel Library Quarterly, 14*, (40-56).

Robertson, S.E. (1981). Term frequency and term value. In: *Theoretical Issues in Information Retrieval: Proceedings of the Fourth International Conference on Information Storage and Retrieval, May 31-June 2, 1981, Oakland, CA* (pp. 22-29).

Robertson, S.E. (1990). On term selection for query expansion. *Journal of Documentation, 46*, 359-364.

Robertson, S.E. (1994). Query-document symmetry and dual models. *Journal of documentation, 50*, 233-238.

Robertson, S.E. (Ed.) (1997). Special issue on Okapi and information retrieval research. *Journal of Documentation, 53*.

Robertson, S.E., & Belkin N.J. (1978a). Ranking in principle. *Journal of Documentation, 34*, 2, 93-100.

Robertson, S.E., Belkin N.J. (1978b). Letters to the editor: Ranking in principle. *Journal of Documentation, 34*, p. 351.

Robertson, S.E., & Hancock-Beaulieu, M.M. (1992). On the evaluation of IR systems. *Information Processing and Management, 28*, 457-466.

Robertson, S.E., Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, May-june*, 129-146.

Robertson, S.E. & Thompson, C.L. (1990). Weighted searching: the CIRT experiment. In: *Informatics 10: prospects for intelligent retrieval, Cambridge, 21-23 March, 1989* (pp. 153-166). London: ASLIB.

Robertson, S.E., & Walker, S. (1997). On relevance weights with little relevance information. In: *Proceedings of the Twentieth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, Pennsylvania, 1997* (pp. 16-24).

Robertson, S.E., Bovey, J.D., Thompson, C.L., & Macaskill, M.J. (1986). Weighting, ranking and relevance feedback in a front-end systems. *Journal of Information Science, 12*, 71-75.

Robertson, S.E., Maron, M.E. & Cooper, W.S. (1982). Probability of relevance: A unification of two competing models for document retrieval. *Information Technology, research and Development, 1*, 1-21.

Robertson, S.E., Maron, M.E., & Cooper, W.S. (1983). The unified probabilistic model for IR. In: G. Salton & H.J. Schneider (Eds.), *Proceedings of the 1982 Berlin Conference, Research and Development in Information retrieval* (pp. 108-117). Berlin: Springer-Verlag.

Robertson, S.E., Walker, S., Beaulieu, M.M., Gatford, M., & Payne, A. (1996). Okapi at TREC-4. In: *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*. NIST Special Publication 500-236.

Robins, R.H. (1980). *General linguistics - an introductory survey.* London: Longman.

Ruben D.B. (1992). The communication-information relationship in system-theoretic perspective. *Journal of the American Society for Information Science, 43,* 15-27.

Salton, G. (1989). *Automatic text processing.* Reading, MA: Addison-Wesley.

Salton, G. (1972). Automatic thesaurus construction for information retrieval. *Information processing, 71,* 115-123.

Salton, G. (Ed.) (1971). *The Smart retrieval system: Experiments in automatic document processing.* Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. & McGill, J. (1983). *Introduction to Modern Information Retrieval.* NY: McGraw Hill Inc.

Salton, G., & Wong, A. (1978). Generation and search of clustered files. *ACM Transactions on Data Base Systems, 3,* 321-346.

Salton G., Bucklet, C., & Smith, M. (1990). On the application of syntactic methodologies in automatic text analysis. *Information Processing & Management, 26,* 73-92.

Salton, G., Wu, H., & Yu, C.T. (1981). The measurement of term importance in automatic indexing. *Journal of the American Society for Information Science, 32,* 175-186.

Saracevic, T. (1970). The concept of relevance in information science: a historical review. In: T. Saracevic (Ed.), *Introduction to Information Science* (111-151). New York: R.R. Bowker

Saracevic, T. (1971). Selected results from an inquiry into testing of information retrieval systems. *Journal of the American Society for Information Science, 23,* 126-139.

Saracevic, T (1975). Relevance a review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26,* 321-343.

Saracevic, T. & Kantor, P. (1988). A study of information seeking and retrieving III: Searchers, searches and overlap. *Journal of the American Society for Information Science, 39,* 197-216.

Saussure, F. de (1974). *Course in general linguistics.* London: Fontana.

Schamber, L., Eisenberg, M., & Nilan, M. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management, 26,* 755-776.

Schank, R. (1975). *Conceptual information processing.* New York: North Holland.

Schleifer, R. (1987). *A.J. Greimas and the nature of meaning -linguistics, semiotics, and discourse theory.* London: Croom Helm.

Searle, J.R. (1969). *Speech acts: an essay in the philosophy of language.* Cambridge: Cambridge University Press.

Searle, J.R. (1984). *Minds, brains and science.* London: British Broadcasting Corporation.

Sebeok, T.A. (1976). *Contributions to the doctrine of signs.* New York: University Press of America.

Shoval, P. (1985). Principles, procedures and rules in an expert system for information retrieval. *Information Processing & Management, 21,* 475-487.

Sless, D. (1986). *In search of semiotics.* London: Croom Helm.

Smeaton, A.F. (1991). Prospects for intelligent, language-based information retrieval. *Online Review, 15,* 373-382.

Smeaton, A.F. & van Rijsbergen, C.J. (1983). The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal, 26,* 239-246.

Smith, M.P., Politt, A.S., & Li, C.S. (1992). Evaluation of concept translation through menu navigation in the MenUSE intermediary system. In: T. McEnry & C. Paice (Eds.) *Proceedings of the 14th BCS IRSG Research Colloquium on Information Retrieval* (pp. 38-54). NY: Springer-Verlag.

Sparck Jones, K. (Ed.) (1971). *Automatic keyword classification for information retrieval.* London: Butterworths.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28,* 11-21.

Sparck Jones, K. (1973). Does indexing exhaustivity matter? *Journal of the American Society for Information Science, 24,* 313-316.

Sparck-Jones, K. (Ed.) (1981). *Information retrieval experiment.* London: Butterworths & Co.

Sparck Jones, K., & Jackson, D.M. (1970). The use of automatically-obtained keyword classifications for information retrieval. *Information Storage and retrieval, 5,* 175-201.

Swanson, D. (1987). Two medical literatures that are logically but not biographically connected. *Journal of the American Society for Information Science, 38,* 228-233.

Swanson, D. (1989). Online search for logically-related noninteractive medical literatures: a systematic trial-and-error strategy. *Journal of the American Society for Information Science, 40,* 356-358.

Tague, J.M. (1981). The pragmatics of information retrieval experimentation. In: K. Sparck Jones (Ed.), *Information Retrieval Experiment.* London: Butterworths.

Thomas, J.-J. (1993). Texts on-line. *Computer and the Humanities, 27,* 93-104.

Turtle, H., & Croft, W.B. (1990). Inference networks for document retrieval. In: *Proceedings of the Thirteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Brussels, Belgium, Sept. 5-7, 1989* (pp. 1-24).

Vakkari, P., & Cronin, B. (Ed.) (1992). *Conceptions of Library and Information Science - Historical, Empirical and Theoretical Perspectives: Proceedings of the International Conference Held for the Celebration of the 20th Anniversary of the Department of Information Studies, University of Tampera, Finland, 26-28 August 1991.* London: Taylor Graham.

van Rijsbergen, C.J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation, 33,* 106-119.

van Rijsbergen, C.J. (1989). Towards an information logic. In: *Proceedings of the Thirteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Brussels, Belgium, Sept. 5-7, 1989* (pp. 77-86).

van Rijsbergen, C.J., & Croft, W.B. (1975). Document clustering: an evaluation of some experiments with the Cranfield 1400 collection. *Information Processing and Management, 11,* 171-182.

van Rijsbergen, C.J., & Sparck Jones, K. (1973). A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of documentation, 29,* 251-257.

Vickery, B., & Vickery, A. (1992). An application of the language processing to search interface. *Journal of Documentation, 48,* 255-275.

Vickery, B., & Vickery, A. (1993). Online search interface design. *Journal of Documentation, 49,* 103-187.

Wade, S.J., & Willet, P. (1988). INSTRUCT: a teaching package for experimental methods in information retrieval. Part III. Browsing, clustering and query expansion. *Program, 22,* 44-61.

Wade. S.J., Willet, P., & Bawden, D. (1989). SIBRIS: the sandwich interactive browsing and ranking information system. *Journal of the Information Science Principles Practice, 15,* 249-260.

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

Warner, J. (1991). A note on the literal intelligence of computers and documents. *Journal of Documentation, 47,* 167-190.

Warner, J. (1994). *From writing to computers.* London: Routledge.

Weiss, E.C. (1977). *The many faces of information science.* Baulder, Colo.: Westview Press.

Wersig, G. (1993). Information science: the study of postmodern knowledge usage. *Information Processing & Retrieval, 29,* 229-239.

Wiesner, S.J. (1988). Two & two a high level system for retrieving related pairs of documents. *SIGOIS-Bulletin, 9,* 1-4.

Wilden, A. (1977). *System and Structure -essays in communication and exchange.* London: Tavistock Publications.

Wilson, P. (1973). Situational relevance. *Information Processing and Storage, 9*, 457-471.

Winograd, T., & Flores, F. (1986). *Understanding computers and cognition - a new foundation for design.* Ablex Publishing Corporation.

Wittgenstein, L. (1958). *Philosophical investigations.* Oxford: Blackwell.

Yip, M.K. (1981). *An expert system for document retrieval.* M.S. thesis, MIT, Cambridge, Mass.

Zunde, P., & Dexter, M.E. (1969). Indexing consistency and quality. *American Documentation, 20*, 259-267.