Gilles Deschâtelets
Université de Montréal. Ecole de Bibliothéconomie et des Sciences de l'Information

The Three Languages Theory in Information Retrieval*

Deschâtelets, G.: **The three languages theory in information retrieval.**
Int. Classif. 13 (1986) No. 3, p. 126–132, 15 refs.

To an overwhelming extent, storage and retrieval systems were designed for information intermediaries who were specialists in formal, controlled documentation languages (e.g. classification systems, indexing languages) and who were then trained to utilize the query language of each retrieval system. However, with the advent of the microcomputer, there now exists, in the information retrieval industry, an obvious will to tackle both the professional and the personal information markets, as evidences by their more sophisticated yet more user-friendly systems and by the design and marketing of all sorts of interface software (front-end, gateway, intermediary). In order to take full advantage of these systems, the user must be able to master three different languages: the natural language of the discipline, the indexing language, and the system's query language. The author defines and characterizes each of these languages and identifies their issues and trends in the IR cycle and specifically in public online search services. Finally he proposes a theoretical model for the analysis of IR languages and suggests a few research avenues. (Author)

1. Introduction

This paper deals with languages, linguistics, classification and indexing from a retrieval perspective. That is, from the point of view of the user, the "end-user", who – if predictions materialize – will soon be or is already sitting in front of his micro, in his home or his office, trying to understand its operating system, its modem and communication software; dialing to a host system, that is a "supermarket" of databases; struggling with a rigid and totally esoteric logon protocol; asking for databases which he knows only from a three-line description in a catalog; using "unnatural" commands and mnemonics, syntactic relationships expressed in terms of boolean logic; using words or phrases which are considered by the system as keywords, descriptors or identifiers in a straightforward character matching process, all the while thinking in more than one dimension, in terms of concepts or ideas; retrieving – if any – one or a few relevant citations and being convinced that they represent 100% recall; having to use still rather strange commands to see or print or display or type or visualize them, to realize that the database only has citations and not the full text of the documents; then either having to go to his library to obtain those documents (and I will avoid any unpleasant remark about that process) or ordering directly online copies of the documents which he will be sent through the mail (and I will again avoid any unpleasant remarks about that process) a few weeks later and for which he will be grossly overcharged, only to find out that the documents do not contain the data or piece of information he was looking for! All this, on a well-designed, ergonomic keyboard . . . probably using one or two fingers.

Surely, there must be easier ways of finding textual information. But things are getting even worse because now, we have just entered the "Era of the End-user", the "ultimate" user, and there is a whole information industry being built around this "person", complete with front-ends, gateways, intermediary software, downloading and post processing facilities and other user-friendly, cordial, convivial devices . . . which, of course, we have to pay for. In all honesty, the online industry has indeed recognized that the whole search process is a mess, that it looks like a 5.000-pieces puzzle for which you would have lost the box and the model picture. Unfortunately, its solutions, so far, have only been to increase the number of pieces in the puzzle . . .

2. The online Search Process

Basically, an *online search* aims at providing a "user", i.e. a person with an information problem, with documents or references to documents that contain an answer to his question or a solution to his problem. Obviously, the user knows or expects the solution to his problem to be found in a document, otherwise he will enquire elsewhere for an answer (colleague, professional). The search is performed on local terminals or microcomputers connected to a host system computer through one or more communication networks. Documents or references to documents (citations) are stored in databases which, in turn, are stored on the host system computer. The search can be done by the user himself or by a search intermediary, an expert searcher acting with or on behalf of the user. Whatever method is chosen, an online search usually consists of a series of steps, decisions and actions (Fig. 1).

3. Problems of Online Searching

A person going online to find a set of documents in answer to a problem (whether a user or an intermediary searching on behalf of a user) is faced with many potential ambushes:

- problems with equipment and software (operating system, modem, communications, printer)
- problems with connection (selection of network, dialing, logging in)
- problems with the selection of appropriate databases
- problems with the host system commands, messages and procedures (query language)
- problems with mapping of the search strategy (concepts, descriptors, keywords, access points, logic, limitations)
- problems with the structure and indexing policy of each databases; problems with the vocabulary and syntax of each database (indexing language)
- problems with the terminology of the domain (jargon)
- problems with the ordering of documents or accessing the full-text of documents
- problems with the creation of personal files.

[*] Based on a paper presented at the 3rd Regional Conference of FID/CR, Montréal, Canada, Sept. 13, 1986.

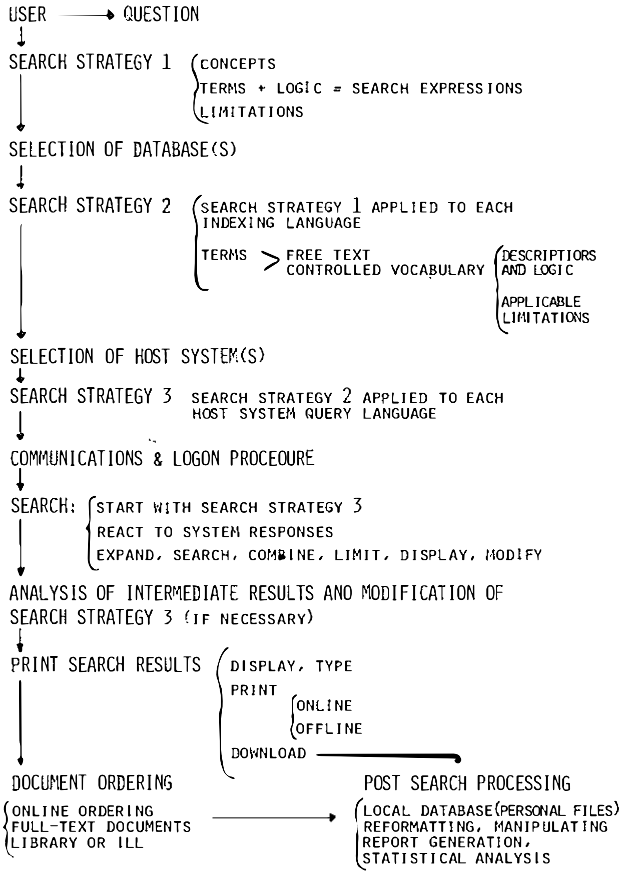

USER
QUESTION
SEARCH STRATEGY 1
CONCEPTS
TERMS + LOGIC = SEARCH EXPRESSIONS
LIMITATIONS
SELECTION OF DATABASE(S)
SEARCH STRATEGY 2
SEARCH STRATEGY 1 APPLIED TO EACH INDEXING LANGUAGE
TERMS
FREE TEXT
CONTROLLED VOCABULARY
DESCRIPTIORS AND LOGIC
APPLICABLE LIMITATIONS
SELECTION OF HOST SYSTEM(S)
SEARCH STRATEGY 3
SEARCH STRATEGY 2 APPLIED TO EACH HOST SYSTEM QUERY LANGUAGE
COMMUNICATIONS & LOGON PROCEDURE
SEARCH:
START WITH SEARCH STRATEGY 3
REACT TO SYSTEM RESPONSES
EXPAND, SEARCH, COMBINE, LIMIT, DISPLAY, MODIFY
ANALYSIS OF INTERMEDIATE RESULTS AND MODIFICATION OF SEARCH STRATEGY 3 (IF NECESSARY)
PRINT SEARCH RESULTS
DISPLAY, TYPE
PRINT
ONLINE
OFFLINE
DOWNLOAD
DOCUMENT ORDERING
ONLINE ORDERING
FULL-TEXT DOCUMENTS
LIBRARY OR ILL
POST SEARCH PROCESSING
LOCAL DATABASE(PERSONAL FILES)
REFORMATTING, MANIPULATING
REPORT GENERATION,
STATISTICAL ANALYSIS

Fig. 1: Online Search Process

Most of these problems can be easily overcome by training and experience. However, they are also greatly amplified considering that:

- there are numerous types of equipment and software; criteria for selecting the best equipment for online searching are not necessarily those of other applications (e.g. word processing, file creation);
- there are many communications networks and each host system has one or more addresses on one or more of these networks (Datapac, Telenet, Tymnet); these networks are interconnected from one country to another;
- as of July 1986, there were 3169 different databases (Cuadra, 1986) commercially accessible on host systems; created by 1494 different producers, these databases are either bibliographic, referral, factual, encyclopedic, numeric or full-text; each database covers one or more domains (chemistry, history, horses, welding), certain types of documents (periodical articles, monographs, reports, theses, conference proceedings); they go back in time to various dates; each database has a unique indexing policy and indexing language (indexing terms controlled to various degrees) which not only vary from one database to another but which can also vary in each database over time;
- as of July 1986, there were 486 different host systems commercially accessible over the world (Cuadra, 1986); each system has its own unique commands, procedures and search facilities (query language); again, these can vary considerably from one system to another and they can vary in each system over time (e.g. Dialog 2, Questel Plus); – "natural language", especially technical or domain-related jargons, can sometimes be far from natural and difficult to understand, even for the specialist.

4. The Three Languages in Online Searching

Any online search, with the present generation of retrieval systems, necessarily makes use of at least three different languages (Fig. 2):

the natural language of the document;
the indexing language of the database; and
the query language of the host system.

According to Webster, a *language* is "a system of communication between humans through written and vocal symbols". Most of our information and documentation systems at present are based on written languages. However, this is likely to change in the near future with the rapid evolution of audio and video storage and retrieval devices (CD-ROMs, videodiscs). Nevertheless, the present discussion is based on written or textual languages.

NL	USER QUESTION	
NL	SEARCH STRATEGY 1 (NATURAL LANGUAGE)	
	SELECTION OF DATABASE(S)	
IL	SEARCH STRATEGY 2 (LANGUAGE OF DATABASES)	
	SELECTION OF HOST SYSTEM(S)	
QL & IL	SEARCH STRATEGY 3 (LANGUAGE OF HOST SYSTEMS)	
QL	COMMUNICATIONS AND LOGON PROCEDURES	
QL & IL	SEARCH	
QL	PRINT SEARCH RESULTS	
QL	DOCUMENT ORDERING	POST-SEARCH PROCESSING (OTHER MACHINE LANGUAGE)

Fig. 2: Online Search Languages

NL	Natural Language
IL	Indexing Language
QL	Query Language

4.1 Natural language

The *natural language* (Fig. 3), in the present context, is "the language we find in documents, without any modification" (Sharp, p. 193). In the online search process, the natural language is that of the authors of documents and that of the users of the systems. It is generally assumed that those are identical or at least compatible. In other words, we assume that the user will "understand" the language of the author of the document. However, there exist numerous linguistic systems (popular language, scholarly language, technical language, expert or specialized language (jargon) and so on). When used to record information in documents, these are all referred to as "natural languages". But this can be very deceiving. Think of the word "information" and its different meanings to a journalist, a politician, a mathematician, a military, a secret agent, a computer specialist, a librarian, an archivist, a linguist or a layman.

One important characteristic of a written or textual language is that it is "information rich". It carries carved and encapsulated messages. It is condensed, compact and generally non redundant. It conveys maximum in-

NATURAL LANGUAGE	INDEXING LANGUAGE	QUERY LANGUAGE
LANGUAGE OF AUTHORS AND USERS	LANGUAGE OF DATABASES (INDEXERS) FOR DESCRIBING THE CONTENT OF DOCUMENTS	LANGUAGE OF HOST SYSTEMS FOR SEARCH APPLICATIONS (ACCESSING DATABASES)
LANGUAGE OF DOCUMENTS NUMEROUS LINGUISTIC SYSTEMS: POPULAR LANGUAGE SCHOLARLY LANGUAGE TECHNICAL LANGUAGE DOMAIN-RELATED LANGUAGE WRITTEN LANGUAGE IS INFORMATION-RICH LEVELS: PHONOLOGICAL MORPHOLOGICAL LEXICAL SYNTACTIC SEMANTIC PRAGMATIC	MORPHOLOGIC FREE VOCABULARY KEYWORDS (IDENTIFIERS) NATURAL LANGUAGE CONTROLLED VOCABULARY SUBJECT HEADINGS (AUTHORITY FILES) THESAURUS SYNTACTIC NATURAL ARTIFICIAL BOOLEAN LOGIC RELATIONAL TREE STRUCTURE MULTICRITERIA ELABORATE STRUCTURE	PROCEDURAL LANGUAGE COMMAND-BASED MORPHOLOGIC PROCEDURES COMMANDS SERVICE SEARCH SEARCH ASSISTANCE EDITION ASSISTANCE KEYWORDS (SPECIAL FUNCTIONS) MESSAGES ANSWER ERROR ASSISTANCE SYNTACTIC SEPARATORS FUNCTION CHARACTERS WRITING RULES QUERY-BASED (MENUS, TREES, SCRIPTS, GRAPHICS, QAS) MORPHOLOGIC PROCEDURES PROMPTS SCRIPTS, MACROS QUESTIONS WINDOWS TAGS MESSAGES ANSWER, ERROR, ASSISTANCE SYNTACTIC

Fig. 3: Languages in Online Searching

formation in a minimum of words. Compare, for instance, 500 words from an ordinary conversation with 500 words in a journal or newspaper article. This characteristic is carried over in indexing languages where subject fields and subfields can be graded from most to less condensed, most to less "information rich":

↑	controlled descriptor or subject heading
	uncontrolled keyword or identifier
level of information "richness"	title word
	abstract word
	full text word

For instance, the word "library" used as a descriptor of a document is more likely to be "descriptive" (or semantically rich) of the content of the document than the same word used in the title, the abstract or somewhere in the text of the document.

4.2 Indexing language

The *indexing language* is the language used to describe the content of a document for classification and retrieval purposes. Indexing can range from the use of a few keywords to all significant words of the title, the abstract or even the complete text of the document. The indexing language can be "natural" or it can be controlled. *Natural language indexing* simply consists of listing all significant words (free vocabulary), as they appear in a document or citation (record), in an online dictionary with pointers to the source record (inverted index). This process can be easily automated. Natural language retrieval or "free-text" searching, thus, refers to a mode of searching where all significant words in a stored document or citation (record) may be used as retrieval keywords.

One very important remark must be made about natural language indexing and searching. However "natural", i.e. close to the document's language, an indexing language may be, there remain two fundamental distinctions between the two. First, their basic objectives are different: while the objective of the author's natural language is to *communicate* ideas to colleagues or potential readers, that of the natural indexing language is to *store and retrieve* documents. Second, the basic attributes of natural language are severely limited by the present machine storage and retrieval techniques. For example, machines cannot recognize other linguistic forms than strings of characters separated by blanks or special codes. The retrieval process then consists in the matching of significant elements in the search question with those already stored in the system. But in this process, natural language looses its multidimensionality, its substance and becomes unidimensional, linear and static. It simply becomes a storage support instead of a communication tool. It becomes artificial as all indexing languages.

Controlled indexing vocabulary is a standardized list of subject terms (descriptors) used in indexing documents. Because of the simple matching process, obviously the same descriptors must also be used for retrieval. These lists of controlled subject terms are sometimes called authority files or thesauri. Thesauri go beyond the simple listing of preferred terms and include a rudimentary form of syntactic relationships between terms (hierarchy, relatedness, synonymy). Retrieval with controlled vocabulary requires the user to know the language, i.e. the appropriate descriptors, their form and the structure of their relationship. Hence he has to learn another language or use vocabulary aids (dictionaries, thesauri, lists of subject headings). However, a good controlled vocabulary will establish a network of cross-references from natural language terms to preferred descriptors.

As any language, indexing and retrieval languages should include two basic components: a *morphological component* (vocabulary) and a *syntactical component* (structure). Both could be natural or artificial, i.e. controlled (Sharp, pp. 192–193). For retrieval purposes,

> "If we try to match the vocabulary *and* the structure of natural language by free-text searching of full texts, then in the nature of things we must know what the document says before we retrieve it. What purpose then in retrieving it?" (Sharp, p. 193)

Considering these two components of language, vocabulary and structure, we can propose the following categories of indexing languages:

1. *Combinatory or multicriteria languages* in which the structure or syntax is completely independent of, and external to the vocabulary; for example, systems with inverted indexes and boolean logic;
2. *Hierarchical or tree-structures languages* in which the vocabulary is preorganized or structured in hierarchies or decision trees; examples are thesauri and classifications;

3. *Elaborate languages* which include more detailed syntactical relationships in the structure of their vocabulary; examples are Syntol, Semantic Code, Precis, Farradane's relational system and NEPHIS.

However, "most information retrieval systems now are based on keywords which alone are not sufficient to express the content or meaning of a document" (Goldsmith, p. 7).

4.3 Query language

The *query language* is that of the host system. It is essentially a *procedural* language, i.e. a totally artificial language designed to accomplish specific tasks or procedures such as logging in, requesting a database, searching for terms and expressions, printing information, and so on. It is an application language which tries to emulate either natural language or indexing language. Other more sophisticated query languages make use of knowledge presentation techniques (e.g. frames, scripts, schemas, graphics). Most query languages have a morphological and syntactical component. The morphological aspect of the query language includes procedures and messages. *Procedures* are expressed in terms of *commands* and *keywords* which are sets of instructions – either words or symbols – directing the computer to take some specific action (e.g. select, display, print, limit). Commands are not standardized from one system to another. For example, in order to search for a term, you might have to use the command SELECT, FIND, SEARCH, QUESTION, CHERCHER or simply to type in the search term without any command depending on the host system. Furthermore, each command can have a mnemonic code or a symbol using one, two, three or four characters (e.g. S, F, SA, BAS, STOP). *Messages* are either prompting messages soliciting some action or response from the user, error messages, or help and assistance messages explaining the meaning of a command or procedure. The format of messages is governed by a set of conventions called *protocols*. Messages can be polite ("please login"), friendly ("goodbye!"), straightforward ("syntax error, invalid command format"), obscure ("line interrupt 4B72", "error 58"), esoteric (e.g. the : or ? prompts) or gossipy (e.g. the automatic lengthy news of the system after a logon).

The syntactical aspect of the query language consists of the "grammatical rules" of the language. These are generally more difficult to learn than the commands. Syntactical aspects are much more scattered and independent than morphological aspects in query languages. They include such functions as separators, function characters and writing rules. Every online searcher is familiar with the "nightmare of the blank", i.e. the character used by each system for separating words: a blank, a comma, a period, a semi-colon, a dollar sign, and so on. Function characters are also numerous and unstandardized. For example, truncation symbols vary from system to system: Stairs ($), ORBIT (:), Dialog or ESA/IRS (?), Mistral or Questel (+), BASIS, CAN/OLE, INFOLINE or QL systems (*).

This is not only a syntactical aspect of the query language, it also becomes an ergonomic factor. Query languages show no syntactical rules and their learning and mastering is very difficult, even for the trained searcher. Furthermore, the query language closely complements the indexing language in the search and retrieval process with additional facilities such as boolean logic, word adjacency and proximity, search limitations (fields, dates, special codes), truncation and masking; these could all be construed as "structural" elements of indexing languages of all databases available on *that* system. Hence the query language and the indexing language are sometimes considered as two elements of the global "retrieval language" (Chaumier, p. 68).

Query languages can be command-based or query-based. Query-based languages include: – menus (tree structure), graphics (fame or window structure), scripts (schema or bordereau structure) and question answering (dialogue structure).

> "At present, most host systems are "command-based" rather than "query-based". In a command-based system it is up to the user to initiate the search, give instructions to the system and decide on the next step to be taken. This means that the user has to learn a command language and inevitably a certain amount of training is required. A query-based system is totally different – the system guides the user by prompting and asking questions, giving advice and controlling the overall patterns of events. This approach means that no training is required and the system caters for the new, inexperienced or occasional user". (Goldsmith, p. 7).

Thus, in a *command-based system*, the user has to initiate the dialogue with the system. This supposes that he knows the "language" and the operations that the language will generate. He is in complete charge of the system and hence, in order to be minimally efficient, he must be knowledgeable of both the language and the procedures. He must have a minimum level of expertise. That is one major reason why most end-users – who generally are novice and causal users of online systems – never really took over the online search process.

On the other hand, in *query-based systems*, it is the system which initiates the search and controls the dialog. This can be accomplished by prompting and question-asking or by controlling the overall pattern of events in the search through menus, scripts or graphic hand-holding.

5. The Three Language Theory

This constitutes the core of the "three language theory". The proposition which is more theorematic than theoretical and more empirical than scientific, can be very briefly outlined as follows:

Observation: in order to conduct online searches with a minimum of effectiveness and efficiency, the user must learn and use indexing language(s) and query language(s) in addition to the language of the discipline.

Assumption: the user is already familiar with the language of the discipline, at least enough to read and understand the documents to find the solution to his problem.

Predicate: for the task at hand (finding textual information) it is the language of the discipline which requires the minimum (learning) effort on the part of the user.

Issue: the indexing and query languages in online searching should be as close as possible to the natural language of the discipline.

Solutions: there are three alternative approaches to this problem: the (human) search intermediary, the natural retrieval language, and the intelligent interface.

5.1 The search intermediary solution

The search intermediary solution has, until now, been *the* solution of the online industry: a human search intermediary, an expert, playing the role of the interface between the user and the system/database/document. The expertise of search intermediaries consists of a good knowledge of

- indexing languages and database structure
- available systems and databases
- the query language
- interview and negotiation techniques and of
- equipment and communication procedures

Although knowledge of the natural language of the domain is an asset, it is not compulsory and the search intermediary can (and often) do without it, compensating by other techniques such as having the user present during the search, longer, more careful search preparation, good search interview. All in all, the search intermediary solution remains very acceptable. However, it also presents certain flaws:

a) it creates dependence of users upon intermediaries;
b) it can bring losses of information and misunderstandings between the user and the intermediary;
c) it is generally quite expensive (Deschâtelets, 1983).

5.2 The natural retrieval language solution

A second solution to the problem of the three languages in online searching is to create a *totally unified natural language* for IR systems. This meta-language would incorporate the indexing language of the database and the query language of the host system and would be as close as possible to the natural language of the search domain. It is the ultimate solution.

Obviously, we are still a long way from such a solution which is referred to as natural language processing (NLP). As Doszkocs points out (p. 192):

> "The basic long-term dilemma of researchers in IR has been the problem of dealing with the content of unstructured natural-language document texts in the absence of an adequate unified theory of language and meaning. Investigators have been confronted with the variability of ways in which the same ideas and topics can be expressed by different authors, abstractors, indexers, and searchers, the inevitable limitations of the query-matching procedures and the contextual subjectivity of users' relevance judgements concerning retrieved items. In efforts to transcend the limitations of the basic keyword/subject heading/inverted file/Boolean logic search paradigm characteristics of the mechanized systems of the 1960's and early 1970's, IR researchers have come to recognize the inherently probabilistic nature of the information retrieval process.
>
> Linguistic approaches to natural-language processing have played a relatively minor and controversial role in IR research. Many experimental results in fact indicate that the full scope oflanguage understanding may not be needed in IR to achieve acceptable levels of performance, especially when searching text surrogates from which users by definition retrieve not the soughtafter information itself, but merely meaningful pointers to where the actual information may be found".

Applications of NLP in IR fall in one of two categories:
a) for *storage*, NLP has a potential for structuring large bodies of textual information in order to facilitate retrieval of data, facts, units of information, and so on;
b) for *retrieval*, NLP has a potential for the design of a friendly, flexible interface including the handling of convivial query languages.

One *retrieval* application on NLP is question-answering systems (QAS) (Grishman, p. 291–293). Such systems can be used as natural language interfaces for database retrieval. We can easily appreciate the enormous problems associated with QAS that have to deal with heterogenous user populations, thousands of database structures and hundreds of query languages. Such QAS interfaces, as Grisham points out (p. 293), must be readily transportable to new domains, require a substantial amount of "engineering" and still remain much closer to formal languages than to truly unrestricted natural language.

A storage application of NLP is text analysis, that is the conversion of texts into a form more amenable to processing. Many rules are required to handle large bodies of natural language texts: rules to determine the relationships between sentences and to disambiguate sentences based on prior context; rules to extract information needed by a specific application (e.g. class of queries); and so on.

> "The key to text analysis lies in being able to organize this collection of rules. In order to do so, we must first determine the structure of information in the domain whose texts we are trying to process. By this we mean classifying the objects in the domain (forming "semantic classes"), identifying the basic types of facts may combine to form larger structures (. . .) Once a standard set of structures has been defined, the variation in the text can be reduced by mapping the information in the text into these structures". (Grishman, p. 293)

Thus we can conclude, with Grishman, that "automatic text structuring is still some distance from commercial applications". In order to do so, any NLP solution, including text analysis, will have to handle most levels of natural language (with the possible exception of the phonological level for textual documents), as shown (Fig. 4) (adapted from Doszkocs, p. 194).

6. The Intelligent Interface Solntion

The third solution to the problem of end-user searching is the *intelligent interface* alternative. Before we can design a "totally natural" system, capable of handling requests from any user on any domain of knowledge or application, intermediary solutions are required and are being proposed.

The intelligent interface solution basically consists of software and transparent aids and services that assist the user in the various steps of the online search process, as shown in Fig. 5. We can distinguish four types of intelligent interfaces: front-end, gateway, intermediary and post-processor:

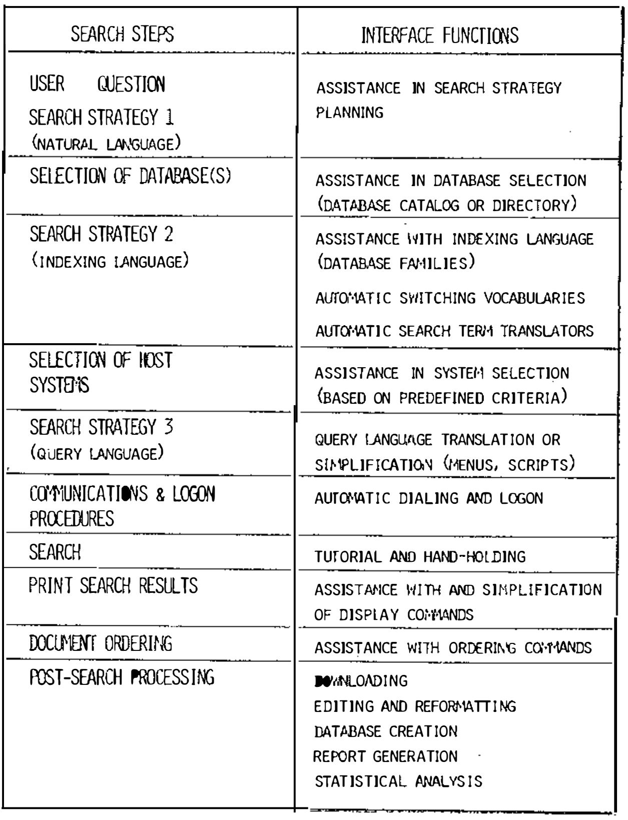
> "The "user-friendly" or "user-cordial" aspect associated with any kind of front end or interface simply indicates that it is easy to use and usually implies easy to learn; it in some way simplifies use and generally substitutes (or reduces the need) for a user's manual or online consultation of documentation. The *"intermediary"* aspect refers to a system that in some way is a surrogate for, or takes the place of, the intermediary searcher. The *"front end"* or "interface" aspect of a system indicates that the system is used in front of, or

between, the user and a target database. The *"postprocessor"* concept associated with a system simply indicates that output from a search is processed in some way that goes beyond the normal processing provided by the online system. The *"gateway"* aspect refers to the ability of one system to provide a pass through another system". (M.E. Williams, 1985, p. 1)

The basic objective of intelligent interfaces is to capture into software the expertise of the search intermediary.

LEVEL	DEFINITION	APPLICATION IN IR
PHONOLOGICAL	Treatment of speech sound	Single or multi-character truncation and masking
MORPHOLOGICAL	Processing of individual wordforms and recognizable portions of words such as prefixes, infixes, suffixes and compound words	* Commands for (neighbor, expand, lexique) * Single or multi-character truncation and masking
LEXICAL	Operations on full words	* Stopword deletion * Automatic search key substitution or augmentation at indexing or search time (table/thesaurus lookups) * Spelling error detection/correction * Handling of acronyms and abbreviations (table lookups)
SYNTACTIC	Identification of structural units, e.g. non phrases	* Not used in IR systems * Quasi-syntactic analysis routines: – subject headings – limiting facilities – adjacency, proximity, and string searching
SEMANTIC	Adding or using contextual knowledge to represent the "meaning" of natural-language texts	* Not used in IR systems * Instead, vocabulary aids are provided as auxiliary searc files or table-lookup and mappingprocedures: – automatic display and use of cross-references, synonyms and related terms from thesauri, subject-headings and classification systems (e.g. MEDLINE's tree structure and "explode" command) – associative term displays (e.g. ESA/IRS's zoom command) – highlighting of matching search terms in display contexts
PRAGMATIC	Uses information about real-life objects and constructs to help in meaning disambiguation	* Not used in IR systems * Manually constructed controlled vocabularies * Cited and citing references, cocitationclusters, dynamic term association displays (indirect methods of information linkage)

Fig. 4: Levels of Language Processing



SEARCH STEPS	INTERFACE FUNCTIONS
USER QUESTION SEARCH STRATEGY 1 (NATURAL LANGUAGE)	ASSISTANCE IN SEARCH STRATEGY PLANNING
SELECTION OF DATABASE(S)	ASSISTANCE IN DATABASE SELECTION (DATABASE CATALOG OR DIRECTORY)
SEARCH STRATEGY 2 (INDEXING LANGUAGE)	ASSISTANCE WITH INDEXING LANGUAGE (DATABASE FAMILIES) AUTOMATIC SWITCHING VOCABULARIES AUTOMATIC SEARCH TERM TRANSLATORS
SELECTION OF HOST SYSTEMS	ASSISTANCE IN SYSTEM SELECTION (BASED ON PREDEFINED CRITERIA)
SEARCH STRATEGY 3 (QUERY LANGUAGE)	QUERY LANGUAGE TRANSLATION OR SIMPLIFICATION (MENUS, SCRIPTS)
COMMUNICATIONS & LOGON PROCEDURES	AUTOMATIC DIALING AND LOGON
SEARCH	TUTORIAL AND HAND-HOLDING
PRINT SEARCH RESULTS	ASSISTANCE WITH AND SIMPLIFICATION OF DISPLAY COMMANDS
DOCUMENT ORDERING	ASSISTANCE WITH ORDERING COMMANDS
POST-SEARCH PROCESSING	DOWNLOADING EDITING AND REFORMATTING DATABASE CREATION REPORT GENERATION STATISTICAL ANALYSIS

Fig. 5: Functions of Intelligent Interfaces

This expertise includes a series of retrieval-related activites (conversion, routing, selection, evaluation) (M.E. Williams, 1986, pp. 207–209). It can also be expressed in terms of the online "behavior" of the search intermediary.

This searching behavior of human intermediaries has been analyzed by Fidel (1985; 1986) in terms of "moves" and decision trees. She found a routine for the selection of search keys: "The decision routine clearly shows that the process of selecting search keys as performed by online searchers can be formalized into a decision tree". (Fidel, 1986, p. 42). A complete set of formal rules for the selection of search keys could thus be identified and "automated to significantly enhance the adaptability of intermediary expert systems" (Fidel, 1986, p. 37).

Over 50 "intelligent" interface products are now commercially available (eg. Sci-Mate, Pro-Search, Search Master, Search Helper, Easy Net, etc.). None includes all of the activities and behavior of the search intermediary. In fact, very few go much further than a few converting and routing activities. No automated selection or evaluation feature yet exists in any commercial product.

Although many transparent search assistance features have been adopted by commercial services and embodied into commercial products, obviously we are still a long way from the total intelligent interface, the "one-stop" searching tool. Indeed, no product can pretend to fit all situations and all clienteles. The interface market is likely to specialize. However, as M.E. Williams observes (1986, p. 213):

> "There is no dearth of entrepreneurs producing packages and services to simplify online information retrieval".

She predicts that by 1990, about 85% of the online functions will be automated and available either as products or as services.

7. Conclusion

In this paper, I have tried to analyze the online search process through the issue of the search language or I should say, languages. Indeed, to accomplish the simple task of finding an answer to a question in a document, searchers must learn and work more or less artificial procedural languages (query and indexing) in addition to the jargon of the domain. This represents an enormous obstacle to the nonmediated access of users to online systems and databases.

Solutions to this problem range from natural language processing applications to the design of intelligent interface software acting as transparent intermediary assistants or experts between the user and the system. Although commercially-available products do not yet exhibit much "intelligence" or expertise, research in this area points at interesting developments for the years to come. One such path is intermediary expert systems designed to mediate between end-users and online systems. These expert systems are to act as skilled consultants, incorporating the "expertise" of search intermediaries. This expertise includes: 1) knowledge of the database structure and indexing language, 2) knowledge of the host system query language, 3) knowledge of formal knowledge representation and search strategy preparation, and 4) knowledge of online behavior, that is, online "heuristics". Also, a user-friendly interface is important if the expert system is to act as a skilled consultant.

Very few intermediary expert systems exist on a commercial basis: IT or USERLINK (P.W. Williams, 1985; Goldsmith, 1986), DIALECT (Bassano, 1986), EXPERT (Marcus, 1983), CANSEARCH (Pollitt, 1984). Most of these systems focus on the system side of the interaction, that is, indexing and query language, communications, rather than on the human side, that is, the searcher's behavior. They try to deal with hermeneutics rather than heuristics. As Fidel mentions:

> ". . . since most of these expert systems are based on text analysis rather than on models of human searching, they cannot process request-related criteria, such as precision or recall requirements".
> (Fidel, 1986, p. 37)

On the basis of almost a quarter of a century of online searching, considering the products already available on the market, and observing the present trends of research and development in the field, what can we predict about the future of databases and online searching, especially end-user searching? According to Neufeld (1986, p. 188), in the short term, we could expect the following trends:

1. Information systems are evolving slowly in the direction of more electronic distribution;
2. more users will be searching online;
3. source databases (full-text and numeric) will increase;
4. software and systems will be developed to permit more fact or "knowledge" retrieval;
5. more transparency aids and user-friendly systems will be developed (probably in the form of expert systems);
6. primary and secondary publication will be integrated electronically throughout creation (by authors), production, and distribution, possibly as "hybrid files";
7. databases will be distributed in forms other than magnetic tapes and will include audio and video information as well as text: floppy disc, videodisc, CD-ROM, and compact disc technologies.

Obviously, in the long term, no one can tell if these trends will continue indefinitely. However, if there is to be a migration from search intermediary to end-user searching, the online search process will have to be drastically simplified. Query languages and indexing languages will have to draw much closer to natural language (even domain-related language), whether through intelligent interfaces or intermediary systems or through natural language processing and expertise incorporated in each online system. Of course, we would not have this problem if systems and databases had been standardized from the beginning . . .

References

(1) Bassano, J.C.: DIALECT, un système expert pour la recherche documentaire. Bulletin du C.I.D. No. 22 (Juin 1986) p. 1–96.

(2) Chaumier, J.: Un obstacle à la communication: les langages d'interrogation dans les systèmes documentaires conversationnels. In: Société Française des Sciences de l'Information et de la Communication, Les obstacles à l'information. Congrès de Bordeaux, 22–24 Mai 1980. Talence: LASIC 1981. p. 67–74.

(3) Cuadra Associates: Directory of online databases. New York, Cuadra and Elsevier, vol. 7, No. 3, July 1986.

(4) Deschâtelets, G.: Le coût-bénéfice des médiateurs dans la recherche bibliographique en-ligne. Revue canadienne des sciences de l'information 8 (1983) p. 39–51.

(5) Doszkocs, T.E.: Natural language processing in information retrieval. J. ASIS 37, No. 4 (1986) p. 191–196.

(6) Fidel, R.: Moves in online searching. Online review 9, No. 1 (1985) p. 61–74.

(7) Fidel, R.: Towards expert systems for the selection of search keys. J. ASIS 37, No. 1 (1986) p. 37–44.

(8) Goldsmith, G., Williams, P.W.: Online searching made simple. London: British Library 1986. (Library and Information Research Report, No. 41), 113 p.

(9) Grishman, R.: Natural Language Processing. J. ASIS 35, No. 5 (1984) p. 291–296.

(10) Marcus, R.S.: An experimental comparison of the effectiveness of computers and humans as search intermediaries. J. ASIS 34, No. 6 (1983) p. 381–404.

(11) Neufeld, M.L., Cornog, M.: Database history: from dinosaurs to compact discs. J. ASIS 37 No. 4 (1986) p. 183–190.

(12) Pollitt, A.S.: A front-end system: an expert system as an online search intermediary. ASLIB Proc. 36, No. 5 (1984) p. 229–234.

(13) Williams, M.E.: Transparent information systems through gateways, front ends, intermediaries and interfaces. J. ASIS 37, No. 4 (1986) p. 204–214.

(14) Williams, M.E.: Highlights of the online database field. Gateways, front ends and intermediary systems In: Proc. 6th National Online Meeting, New York, April 30–May 2, 1985. Medford: Learned Inform. 1985. p. 1–4.

(15) Williams, P.W.: The design of an expert system for access to information In: Proc. 9th Int. Online Inform. Meeting, London, 3–5 December, 1985. Oxford and New Jersey: Learned Information 1985. p. 23–29.

Prof. Dr. G. Deschâtelets.
Université de Montréal. Ecole de Bibliothéconomie et
des Sciences de l'Information. CP 6128, Succ. "A"
Montréal H3C 3J7, Canada.