The Complementarity of Natural and Index Language in the Field of Information Supply
An overview of their specific capabilities and limitations

Robert Fugmann

Dr. Robert Fugmann, born in 1927, is a chemist and information scientist. He has been active in the development of the theory and practice of indexing and information supply for 45 years. He has served as ISKO Vice President. Visiting Professor and lecturer at various universities and other institutions in Germany and in the United States. He is author of 5 books, 109 articles, and 13 book reviews and is recipient of awards from FID, the Association of German Chemists, American Society for Information Science, and the American Chemical Society.

R. Fugmann (2002). **The Complementarity of Natural and Index Language in the Field of Information Supply. An overview of their specific capabilities and limitations.** *Knowledge Organization, 29(3/4).* 217-230. 28 refs.

ABSTRACT: Natural text phrasing is an indeterminate process and, thus, inherently lacks representational predictability. This holds true in particular in the case of general concepts and of their syntactical connectivity. Hence, natural language query phrasing and searching is an unending adventure of trial and error and, in most cases, has an unsatisfactory outcome with respect to the recall and precision ratios of the responses. Human indexing is based on knowledgeable document interpretation and aims – among other things – at introducing predictability into the representation of documents. Due to the indeterminacy of natural language text phrasing and image construction, any adequate indexing is also indeterminate in nature and therefore inherently defies any satisfactory algorithmization. But human indexing suffers from a different set of deficiencies which are absent in the processing of non-interpreted natural language. An optimally effective information system combines both types of language in such a manner that their specific strengths are preserved and their weaknesses are avoided. If the goal is a large and enduring information system for more than merely known-item searches, the expenditure for an advanced index language and its knowledgeable and careful employment is unavoidable.

1 Introduction

For decades the capabilities and limitations of natural and index languages have constituted a highly controversial topic in information theory and practice. The superiority of each type of language for its employment in information systems, or at least their equivalence has often been stated, and the advocates of both opposite approaches have been able to submit

seemingly convincing examples for the proof of their stance.

In a way the situation resembles the case of Galileo who made the statement that all bodies fall at the same speed. This was in opposition to the dominant view, which maintained that the lighter a body the slower it would fall. The dispute was settled when the vacuum was detected later. The conflict had been caused only through disregarding air resistance, which was most evident only in the case of the light bodies.

If we review the voluminous literature on the controversy of natural versus index language, we will also realize that several factors have been disregarded although they have already been extensively discussed. If they had only been better taken into consideration, the controversy could have been reconciled early, much to the advantage of the scientific community. The capabilities and limitations of both approaches would have been better understood and we would not have pursued an EITHER – OR approach but one of BOTH – AND, that is, an approach in which both types of language play their adequate roles. An overview of a possible interplay of both types of language is depicted in Figure 1.

2 The peculiarity of the retrieval situation

First of all, we must refute an argument put forward in favor of natural language information systems. It has often been argued that natural language is obvious for the retrieval purpose *because* it is well known to everybody and has proved so highly effective in all of our human-to-human communication. The validity of this argument is highly questionable, as the following consideration will demonstrate.

In the process of *inter-human communication*, for which natural language has developed, we perceive signs (word, images) and we infer the meaning and the concepts intended to be conveyed. This process is typically one of the *a posteriori* type (see Figure 1, p. 219).

In *retrieval*, however, we encounter the opposite direction of inference. Here, meaning and concept are given, that is, what is of interest to the questioner. What is to be contrived and inferred are the *expressions* through which the requested concept might have been represented in the search file. These expressions will constitute the search parameters in a computerized file or in a book index. This is an inference of the *a priori type*, that is, one much harder to execute than in the case of the *a posteriori* type. It requires the existence and the knowledge of the regularity of the process at hand. For example, it is much easier to *describe* a prevailing weather situation than to *predict* a future one.

Hence, natural language has developed for a purpose inherently different from that of retrieval, and it is far from obvious that this language should *necessarily* be suitable for retrieval also.

For a more detailed discussion of this topic we will define some concepts, which are assumed to be helpful for the clarification of the language issue.

3 Concept

In the framework of our analysis of the strengths and weaknesses of both types of language, we use a classical definition of "concept" as the imagined sum of the essential and true statements that can be made on an item of reference. Each of these statements constitutes one of the various *conceptual features* of a concept out of which the concept is composed, at least in the view of the particular subject field.[1]

For good reasons,[2] we refrain from regarding "concept" as the meaning of a word – a view which is widely encountered (and justified) in linguistics. With an eye to the ease of processing concepts for retrieval, we draw a line between *individual concepts* and *general concepts*. In our context an individual concept is one to which no meaningful conceptual feature can be added.[3]

Examples are persons, institutions, rivers, mountains, chemical elements and substances, for example, iron, acetic acid, and so forth.

The individual concept is completely defined through its sum of conceptual features. Each of its conceptual features is known or can be ascertained and, in case of demand, the less essential ones included. For example, individual concepts can be arranged according to *any* of their features, because these features are known in their entirety or can be looked up. This fact strongly facilitates the processing of individual concepts, as we will soon see below.

In contrast, a general concept is one to which at least another conceptual feature can be added.

Examples are metals (iron, copper, etc.,), towns (Berlin, Paris, London. etc.,), insects (beetles, butterflies, etc.) and processes such as transportation, swimming, teaching, dressing, dying, flying, and so forth.

LANGUAGE, employed for two different purposes

Purpose	Requirements	Direction of inference	Suitability of Natural language	Suitability of Index-language
	Given: Expression	– *Inferred: Concept*		
Inter-human communi-cation	Interpretability, transparency of presentation	**a posteriori,** i.e. *Perceiving* something existing and processing the perception	yes	no
		Reversely:		
Retrieval:	*Given: Concept*	– *To be inferred (imagined, guessed): Expression*		
	Predictability of the mode(s) of Expression for the requested concept	**a priori,** i.e. *anticipating* something not having been perceived	only moderately suitable [1]	yes [2),3)]

1) Suitable for
Questions of recall because the mode of expression is known and need not be predicted or anticipated (e.g., known item searches), and for filing expressions which need not or cannot be interpreted.
Moderately suitable for
Individual concepts (as opposed to general concepts) because their modes of expression *are* fairly well predictable or can be looked up.
Unsuitable for
Questions of discovery, general concepts, concept connectivities because the natural language expressions *are* not sufficiently well predictable here. –

2) Well suitable under the proviso that
a) the text and images are carefully and knowledgably interpreted by an indexer (whereby the meaning of the natural language expression is recognized),
b) the essence of the documents to be selected for the search file is recognized on the basis of those conceptual – categories which have been stated by the system users beforehand,
c) the paraphrasing mode of expression for concepts *are* reliably lexicalized,
d) the ellipses in natural language expressions *are* filled up,
e) the translation of the essence of the documents into the index language is executed according to Cutter's rule (In large information systems, this requires a sound vocabulary structure and an index language grammar in addition to the vocabulary),
f) a good indexing program is used to achieve sufficient degrees of representational predictability and fidelity for that part of the indexing work which is susceptible for mechanization (e.g., sorting, search for the most appropriate terms in the vocabulary) .

The processes a)–e) are indeterminate in nature because of the inherent indeterminacy and unpredictability of natural language expression, especially for general concepts and syntactical concept connectivities. Hence, these) processes defy any satisfactory mechanization. Any program with this goal will inevitably produce many irrelevant responses and much information 1oss, i.e., will work at low ratios of precision and recall.
Paraphrase lexicalization and ellipses filling are crucial for high recall ratios and constitute a source of considerable expenditure in intellectual indexing.

3) Even the best intellectual indexing is bound to suffer from omissions (especially in the peripheral fields of interest) and from lack of recency. These gaps can be mitigated or closed through
g) an additional good automatic indexing and/or
h) an additional good computer-linguistic full text search,

Figure 1. Different purposes of language and the complementaritv of natural and index language

In any processing of general concepts we must concentrate on the *essential* conceptual features because these are the only ones which may be presented in a document or in which a questioner may be interested. This aggravates the processing of general concepts, as we will soon see.

4 Expression

The human always needs some lingual or pictorial mode of expression in order to convey ideas and feelings. In our context, we need not study the various peculiarities of spoken and written language in which the linguist must be interested. Rather (and pragmatically), we distinguish the *lexical expression* from the *non-lexical* (i.e., paraphrasing definition-like) *expression*[4] because they display a crucial difference with respect to the *ambiguity* and *multiplicity* in which they are encountered for the representation of a concept.

As far as *expression ambiguity* is concerned, it often constitutes an obstacle to sufficiently precise searches. An ambiguous expression, phrased as a search parameter, will necessarily retrieve responses with different meanings of the search term, and only one of them will be the one in which the searcher is interested.

As far as a *multiplicity of expressions* where the concept of interest is concerned, it will constitute an insurmountable obstacle to sufficiently complete searches, if this multiplicity is an infinite one. It is impossible to compile an infinite number of alternative search parameters, each of them tailored to one of these possibilities.

5 The dependence of the mode of expression on the type of concept

Generally, knowing which of both modes of expression is preferred in our common communication depends on the type of concepts. By tacit convention and for the sake of comfort, *individual concepts* are almost exclusively expressed through *lexical expressions* of which only a single one or only a small number is in use. These expressions are remembered in cases of "questions of recall" (Bernier 1960; see also section 7)[5] or can be looked up in dictionaries. Hence, and for the purpose of retrieval, these lexical expressions can almost completely be compiled for a concept under consideration. Hence, and most important for retrieval, the modes of expression for *individual concepts* in natural language are fairly well predictable. Consequently, searches for an individual concept can fairly easily be made near to complete in natural language files.

In contrast, *general concepts* are often expressed in the non-lexical, paraphrasing, definition-like mode of expression in natural language, that is, in an *infinity* of variations, from which an author has an entirely free choice. In other words, it is *unpredictable* in which modes of expression a general concept may have entered the search file in full text or free text storage (cf. for example, Blair 2002)[6]; Gesellschaft für Klassifikation, 1985). However, expression predictability is an indispensable requirement for the search in an information store (cf. for example, Bates, 1998, pp. 1188, 1202; Fugmann, 1985, p. 121 "axiom of predictability"; Fugmann, 1993, p. 59 ff,[7] see also Figure 1).

Consequently, in case of a search for a general concept in a natural language file, it is inherently impossible to phrase a set of search parameters anything like complete. Correspondingly, any search for a general concept in a natural language search file will inevitably be incomplete.

In the initial state of the occurrence of a general concept, it must always be expressed exclusively in the non-lexical mode. Only when and if the demand for easy communication increases will a neologism be created (cf. Coates, 1960, p. 21).

For example, it was as long as ten years after the observation of the first cases of AIDS that this term was created, and earlier literature proved to be almost irretrievable. In other words, the general concept was *lexicalized* at that point in time, its expressions were made predictable and the concept was thus made accessible for retrieval. But even after its lexicalization a general concept is often (and for good reason) still expressed in the non-lexical mode.

A concept can have many expressions and an expression can mean many concepts. It is for good reason that in Ranganathan's analytico-synthetic approach the verbal plane and the idea plane (i.e., the plane of the concepts) are meticulously kept separate (Ranganathan, S.R.; Gopinath.. M.A., 1967, p. 327 ff.).

If concept and expression are equated, as is presently done in mainstream "modern" information scientific research, there are no prospects of achieving an overview of the problems encountered here.

6 Indeterminacy in natural language processing

It is conducive to the understanding of the capabilities of mechanized natural language processing to distinguish the determinate processes from the indeter-

minate ones. An indeterminate process is one that proceeds in an unpredictable manner, and, correspondingly, the outcome of which is unpredictable, too.

As we have already stated, the human has an infinite number of possible expressions at disposal for expressing an idea and the author makes an idiosyncratic, unpredictable selection from this infinity.

In particular, this holds true for the non-lexical modes of expression. No two humans will describe the same event or idea in the same words. No individual human will, at different points in time, choose the same words for the same event or idea. Let two humans write abstracts of the same original: they will write different abstracts. In spite of the unpredictability and lack of consistency inherent in all of these processes, all these products are valuable. Hence, consistency cannot constitute a reliable criterion for indexing quality (cf. for example, Soergel, 1994, p. 594)[8].

An indeterminate process can proceed in an infinite number of variations. In an attempt to mechanize or simulate its progress, one would have to contrive and take into consideration an infinite number of possible factors exerting an influence on the process. One would have to lay down instructions for the infinity of all these cases in the program *in advance.* An infinite number of programmers would have to work for an infinite span of time to be successful in such an attempt.

Of course it is possible to lay down instructions for a *selection of obvious examples,* a selection that is, however, always only a tiny one in view of the infinity of possible cases the program must be able to manage. In each of the selected cases, the program will *y* work correctly only *deceptively*.

In spite of these considerations, some representatives of "hard" artificial intelligence (or the vendors of their products) claim the ability to mechanize indeterminate processes, too, and in fact not only restricted to an inherently highly incomplete selection of examples. This is another story that is not discussed here.

Since any phrasing of natural language text is an indeterminate process (Figure 1, bottom), any processing of natural language text is inherently indeterminate in nature, also. Consequently, any natural language processing defies *satisfactory* mechanization, that is, it defies a type of mechanization, which goes beyond the processing of only a (necessarily highly incomplete) selection of examples.

But in every advanced information system there are many processes of the *determinate* type where advanced technology can be employed to great advantage, apart from the input and search routines. Examples are: sorting, arranging and re-arranging the vocabulary terms in hierarchies; searching terms in the vocabulary; and so forth.

Here, the use of technology is much in the interest of the ease, speed, thoroughness and quality of the processes to be executed (Figure 1).

7 Recall vs. discovery in the information search

When searching for an object of interest, it makes a great difference whether the object is known to the searcher in detail beforehand ("question of recall," "known item search") or not ("question of discovery" Bernier, 1960, Figure 1 (1)). Here it is even often uncertain whether the object being searched actually exists.

In case of a question of recall, *any detail* of the document to be retrieved can be used as a search parameter, provided the detail is remembered.

The expression is already given prior to the search and can be used as a search parameter without any further interpretation (hoping that the expression is not too ambiguous). Here we need *not predict* how the object of interest may have been expressed in the search file because the expression is known and given. The natural language phrasing of the object of interest in the search file is fairly well suitable for this purpose (see Figure 1 (1)).

On the other hand, and in questions of discovery, we do not know the details of the documents of interest, especially not the expressions that might have been chosen by the authors for the concepts of interest. These questions require the *predictability of the mode of expression of what is of interest to the searcher (see Figure 1), since these expressions are hardly predictable in natural language* questions of discovery require the essence of the documents to be translated into another, necessarily artificial language, that is, into one in which the expressions for concepts are *predictable.* This requires the translation into an "*index language*,"[9] that is, they require indexing (see section 9).

It is true that this requires considerable effort in the input stage. But this effort is unavoidable if questions of discovery are to be addressed (and also questions for general concepts and for concept connectivity), and if tremendous loss of relevant information is to be avoided.

Any adequate translation is preceded by the understanding and *interpretation* of the text. Accordingly, any translation is an indeterminate process because it starts at an indeterminate point of departure. In the following we will look at the various achievements that have to be contributed in the interpretation of texts for adequate information retrieval.

8 Interpretation

Any message, which the human receives, exerts its effect only through *interpretation.* Without appropriate interpretation, any message is meaningless and fails to convey what the sender intends to communicate (see Figure 1). For example, it is not sufficient merely to perceive an electromagnetic wave of a certain frequency and intensity. We must always conclude what the signal means. It is not sufficient to smell a certain flavor. We do not – subconsciously –content ourselves with perceiving a pattern of black and white pixels. Much more, we interpret these signals as the image of a face, a landscape, a building, and so forth. Signal interpretation constitutes an inherent feature of any living entity, animals and even plants (cf. for example, Budd, 1995, p 307). Written, audible or pictorial messages do not constitute an exception to the necessity of interpretation.

Especially when we execute the interpretation of texts, we more or less subconsciously recognize the *meaning of ambiguous words* through realizing the context in which they are embedded. Any adequate indexing of documents comprises (see Figure 1 (2)) the *recognition of the meaning of ambiguous words* through realizing the context in which it is embedded, the *recognition of the essence* of the documents to be made retrievable,[10] the *lexicalization of paraphrases* through their translation into the corresponding (lexical) terms of natural or technical language, at least as far as the appropriate terms are already in existence (it will be just these lexical units with which the searches will later be undertaken). An example is the translation of the text passage "the patient displayed a pathological fear of and opposition to entering any type of boat or other sea vessel" into "thalassophobia." Also, the linguistic *ellipses* (Ranganathan, 1964, pp. 90, 129) *to be filled* up in order to make accessible for retrieval what is only implied in the text (cf. for example, Green, 1991, p. 84[11]; Roberts, 1997[12]; Hjorland, 1997, p. 26[13]). For example, if an index language contains (and offers for retrieval) the geographical term "Antarctic" then a document on penguins on the South Shetland Islands must be made accessible through this term no matter whether or not "Antarctic" occurs in the document.

Since any text and image interpretation is an inherently indeterminate process, any adequate document interpretation cannot adequately be mechanized, as is stated in section 6.

Mechanically executing interpretation on the basis of a (inevitably highly incomplete) set of selected examples for which processing instructions have been laid down in the algorithm does not constitute an adequate solution. It is true for these selected cases – and treacherously enough – the program works to satisfaction. But the program has no instructions for managing the stream of hitherto unknown cases, which will incessantly enter the search file.

Consequently any *autonomous* mechanized text processing for the purpose of indexing (i.e., one which proceeds without any human intervention) will severely lack the achievements contributed through adequate human text interpretation. Low ratios of precision and low ratios of recall are the consequence.

On the other hand, autonomous, mechanized text processing will be fairly successful in cases when interpretation is not necessary or can be dispensed without severe disadvantages. This holds in the cases mentioned (Figure 1 (1); see also section 11).

9 Intellectual indexing

Intellectual indexing of texts (and also of images) is executed in several variations. Presently "indexing" is widely regarded as the mere *extraction* of *text words.* This approach of "word indexing" does not yield genuine indexes but merely concordances, that is, a list of locators for text words.

We base our considerations on a definition of indexing stated by FID/Classification Research according to which "indexing" is the description of the essential contents of a document, by extraction and/or assignment of significant terms with or without syntactical relationships with a sufficient degree of [representational] fidelity and of [representational] predictability for retrieval demands.[14] Hence, indexing comprises two different steps: that of essence recognition and that of essence representation.

Information retrieval has often been misunderstood and merely viewed as the (purely mechanical) process of locating those words with which a questioner has phrased the topic of interest. Word indexing serves this purpose. However, and much more, a searcher is interested in retrieving *what is meant*

through the verbalization of the topic of interest, irrespective of the mode of expression that an author might have chosen to express the idea of interest.

This type of "concept indexing" always includes the transition from the verbal plane to the idea plane, that is, the *step of interpretation* (see section 8), a step which can, through its indeterminate nature, only inherently be executed inadequately in the algorithmic way.

In many variations of indexing the index language vocabulary is used in a manner against which Cutter[15] had advised more than hundred years ago. Cutter's "best fitting indexing" (as we shall call it henceforth) has for good reasons been practiced extensively in the library profession but it has elsewhere fallen into oblivion. Here, one contents oneself with merely "controlled" indexing.

Furthermore, we seldom encounter an index language grammar as a complement to its vocabulary and if we do, it mostly leans on the unpredictability of natural language and is therefore almost without value for retrieval. Semantic vocabulary categorization is seldom encountered these days.[16] This omission renders the indexing procedure highly unreliable. This uncertainty in retrieval is the source of unnecessarily low ratios of precision and recall in searches. For all these reasons and more, contemporary intellectual indexing is far from exploiting its capabilities to the full.

If we are going to assess the capabilities and limitations of intellectual indexing and those of algorithmic natural language text processing for the purpose of indexing, we must distinguish several variations of indexing in different degrees of effectiveness and input expense.

9.1 Extractive indexing

Extractive indexing is the most simplified version of indexing. Text words are extracted and – in more or less modified version – entered in the search file. The only achievement of the indexer in this case is essence selection through the extraction of keywords. No index language vocabulary is being used here let alone an index language grammar. Representational predictability is at its lowest here. Therefore the recall values in retrieval after such a variation of indexing are correspondingly low.

Representational fidelity and indexing specificity are only seemingly high in this approach because the apparently specific terms are presented in low predictability. Therefore, they are of only limited use for adequate retrieval.

Extractive indexing is useful for questions of recall (see Figure 1 (1); see also section 7) because predictability is not demanded here. It is (even if only moderately) suitable for individual concepts taken in the above-defined meaning because these concepts are normally lexicalized. Their various names (e.g. synonyms) can be looked up and used as alternatives in the query.

Extractive indexing is easiest to mechanize: for example, through stopword lists and/or positive lists. It is the cheapest but also least effective variation of indexing because there is no meaning recognition and clarification, no paraphrase lexicalization, no ellipses filling. In short, there is an only highly incomplete document interpretation. It is only with hesitation that we use the term "indexing" here because it does not meet the definition of indexing already presented.

However, extractive indexing (as is the case with free "indexing" also) can complement all variations of intellectual assignment indexing; that is, controlled and best-fitting indexing in quite a particular manner. Here, an indexer is often not sure whether a concept deserves being entered into the indexing *vocabulary. The terms for these concepts can be collected as descriptor candidates.* From time to time this bag of candidates is emptied and the appropriate decisions are made on their future use as descriptors.

9.2 Free indexing

Free indexing is another variation of low *input* expenditure. The vocabulary is not restricted to that of the documents, and assignment indexing is possible here. In practice, however, this possibility is seldom used because it is the goal of free indexing to circumvent the input expenditure incurred through the construction and use of an index language. Hence, free indexing only just exceeds the limited capabilities of extractive indexing because it also lacks the predictability of the expressions for the appropriate concepts.

9.3 Controlled Indexing

In controlled indexing, typically only those terms are *permitted* for indexing which have been compiled in the agreed upon index language vocabulary. The emphasis is on "permitted." Commonly the indexer is not obliged and not prepared to search for and to use the *most appropriate term* available from the vocabulary for the representation of a concept of interest in the search file. As a consequence, representational

predictability is impaired and the *recall ratio* is only mediocre.

If the searcher, knowing this weakness of the indexing procedure, as a precaution includes less appropriate vocabulary terms in the query, it is in order to counteract recall decrease and precision will inevitably decrease: An imprecise query necessarily produces imprecise responses ("axiom of representational fidelity"; cf. Fugmann, 1985, p. 123; 1993, p.65[17]). Controlled indexing requires knowledgeable text *interpretation,* which can, due to its indeterminate nature, be executed satisfactorily only by the expert in the field.

9.4 Best fitting indexing

A great deal of the uncertainty that prevails in controlled indexing can be overcome through the adherence to Cutter's rule. Here only the *most appropriate terms* from the index language are permitted for the representation of a concept of importance. This type of "best-fitting indexing," as we shall call it henceforth, has been common practice in libraries for more than a century but it has fallen into disuse in "modern" indexing and retrieval practices.

Indexing in this manner results in particularly high ratios both of representational predictability and representational fidelity.[18] The consequence is *typically high ratios both of precision and recall*.[19]

But using a vocabulary in accordance to Cutter's rule has a severe implication. Any adequate search in a search file for the *documents* of interest is, even if only latently, preceded here by a search in the vocabulary of the system for those *terms* that most appropriately represent the topic of interest.

However, for any type of searching, the indexer has at disposal only limited resources of time, memory and patience. This holds true not only for searches in a file of documents but also for searches in a *vocabulary of search terms.* A vocabulary which is chaotic and excessively large, will fail to serve its purpose here.

Hence, executing best fitting indexing according to Cutter makes high demands on the perspicuity and transparency of the vocabulary of the system. In essence, and in very large information systems, a sole *vocabulary* of an index language is overburdened with the task of representing the vast multiplicity of concepts to be managed here. It is only through the introduction of a powerful index language *grammar* that the size of a vocabulary can be kept under control and that a sufficiently high degree of perspicuity can be maintained to enable obeying Cutter's rule.

Any assessment of "human indexing," in particular in comparison with automatic indexing, makes sense only if at the same time it is specified *which of the aforementioned types of human indexing* had been executed in the system under evaluation.

10 Keyword searching in full text files

The most simplified approach to natural language searching is to use keywords which may be assumed to have been used by the authors of relevant texts for the topic of interest. Since there is no text interpretation here, there is no expression disambiguation, no essence selection (both a source of low precision ratios). There is no paraphrase lexicalization and no ellipses filling (both a source of low recall ratios).

But in searches for *known items* and, though only with some reservation, in searches for *individual concepts,* the keyword approach may be useful. Here it can close a gap which may well remain even in most careful human indexing (Figure 1, (3)).

For example, in an index language for the field of information science there would hardly occur the term "battleship." But one may remember a document in which a civil aircraft had been shot by a battleship because "the computer" had erroneously identified the aircraft as hostile and as attacking the battleship. The search term "battleship" will retrieve this document, this certainly in contrast to a search in an indexed file.

Indexed files also notoriously lack recency. Keyword searching may well retrieve documents that have not yet passed the time-consuming indexing step (Figure 1, (3)).

Individual concepts (as opposed to general concepts) are normally represented through lexical expressions the predictability of which is relatively high, as has been stated already. The keyword search is also promising here, in particular if the keyword is one of only little ambiguity (Figure 1, (1)).

11 Mechanized natural language processing for information retrieval

Mechanized natural language processing for retrieval has been investigated in an enormous multitude of variations. Almost all contemporary research concentrates on approaches of this type. In order to assess their capabilities and limitations it is not necessary to study their details (which are often not revealed to the user). We can content ourselves with realizing that all of these mechanized approaches dispense with

adequate text interpretation and that they are therefore restricted to the verbal plane because the transition to the idea (concept) plane (cf. Ranganathan and Gopinath 1967) is an inherently interpretative process. Most of these mechanized approaches rely on text word occurrence and co-occurrence statistics, on positive lists[20] and stopword lists.[21]

Hence, these approaches are devoid of (reliable) meaning recognition, essence selection, paraphrase lexicalization, and ellipses filling. Therefore, they are bound to work with low ratios of precision and recall, irrespective of the algorithmic artifices which have been invented for them.[22] The overall ineffeciency of the statistical approach has been stated explicitly by van Rijsbergen (1990, p. 111[23]).

As far as the effectiveness of *search engines* is concerned, they are also necessarily limited to the verbal plane if adequate text interpretation before input is dispensed with.

They are commonly based on word statistics of full texts or abstracts. Each uses its own search algorithm, the details of which are kept secret from the user. Correspondingly, the search responses are often hard to explain and to justify. A user will hardly be satisfied if told that a document of interest had escaped retrieval because the search word for the concept of interest had occurred too often or too rarely in this document, occurred in the file or in the search requests hitherto executed in the search file. Even the most advanced search engines often produce results in quite unacceptable quality (cf. for example, Bloomfield, 2001, p.65[24]).

On the other hand, it is precisely through their specialization on the verbal plane that algorithmic routines may well constitute a useful complement to intellectual indexing. In several types of searches one can dispense with text- and with query-interpretation without incurring severe disadvantages (see Figure 1 (1)). This is the case with known-item searches and, even if only with some reservations, with searches for individual concepts, in particular, if one can put up with incompleteness of the search responses (cf. Figure 1, (1) and (3)).

An index language always forces the indexer to assume a certain textword meaning because the meaning of the terms in the vocabulary is pre-defined. If the meaning of a natural language expression is highly ambiguous or even unknown, a natural language file offers the possibility of storing the expression *without the need of any interpretation* which may be a highly questionable one.

Terminological investigations are another example of the usefulness of known-item searches in non-interpreted text files. Here one is interested in retrieving an *expression of interest* and in just finding out in which meaning the expression has been used in the past.

An information seeker often transforms a search for a *topic* into a search for the *name of an author*, which is remembered as being closely associated with the topic of interest. In other words, a search for a topic is, although with some distortion, transformed into a search for an individual concept. Again, a natural language file is promising for such a search.

Numerous algorithms have been developed in the past for the *book indexing procedure*. But due to their being restricted to the verbal plane and to their lack of adequate text interpretation they have disappointed when used in practice (for criticism, cf. for example, Mulvany and Milstead, 1994), this in sharp contrast to what has been promised by their producers. If the research efforts continue to be concentrated on information technology, there are no prospects of substantial progress (cf. for example, Wellisch, 1992; Shpackov, 1992; Swanson, 1988).

12 The integrated approach

In the foregoing, it is stated that natural language processing and human indexing serve different purposes and retrieval situations. If combined in a well-considered manner both approaches can very effectively complement each other in such a manner that their specific limitations can be eliminated and their specific strengths are preserved.

Here "best fitting indexing" is combined with full text processing strategies, executed in the full texts of the documents or of their abstracts. They do not only serve for keyword searching but they also serve as a source of instant visual information on the responses to a query. If questions of recall dominate in such an information system, the accessibility of full texts is also advisable for these searches. Such an integrated approach cannot be a cheap one in the input stage because it necessitates the cooperation of the knowledgeable and careful expert.

It is true that advanced intellectual indexing is slow and expensive but "It can save the searcher an incredible amount in lost time and missed documents" (Ilmholtz, 1999). If executed carefully and knowledgeably it can provide a level of retrieval quality which is unattainable solely through mechanized natural language text processing. Such an information

system displays a high survival rate under the steadily changing and increasing burdens of the future.

13 A comparison of input and search strategies

If many circles of natural language technology claim their approaches to be equivalent or even superior to the traditional, intellectual strategies, they must tolerate being compared with the traditional approach. Such a comparison must not be restricted to the most primitive variation of intellectual indexing, that of extractive indexing,[25] that is, that of least expenditure and of least effectiveness.

Such a comparison is depicted in Figure 2.

REQUIREMENTS for the supply of information from a mechanized search file	INPUT AND SEARCH STRATEGY: Full text input and search 1)	Indexing, extractive	free	controlled	best fitting 2)	integrated 3)	automatic
1. Meaning disambiguation	–	–	(+)	+	+	+	–
2. Predictability of essence selection	–	–	–	+	+	+	–
3. Predictability of essence representation	–	–	–	(+)	+	+	–
4. Paraphrase lexicalization	–	–	–	+	+	+	–
5. Ellipses filling	–	–	–	+	+	+	–
6. Nontextual information	–	–	(+)	+	+	+	–
7. Synonym control	–	(+)	(+)	+	+	+	–
8. Descriptor and keyword specificity	(+)	(+)	(+)	(+)	(+)	+	(+)
9. Recency	(+)	(+)	(+)	–	–	(+)	(+)
10 Possibilty of retrieval-suitable syntax	(+)	(+)	(+)	+	+	+	(+)
11. Possibility of collecting descriptor candidates	.	.	–	–	–	+	.
12. Availability of terms that escape indexing	+	+	+	–	–	+	+
13. Input and search economics	disputable						
Evaluations (qualitative) in strengths	$2^1/_2$	3	4	8	$8^1/_2$	$11^1/_2$	$2^1/_2$

Legend: "+": strength
"(+)": limited strength (counted half)
"–": weakness
".": not applicable

1) Lowest input expenditure in the search but highly limited in search effectiveness and survival power except in case of Question-S of recall and textword search.
2) Necessary for adequate effectiveness in an intranet project due to its high recall ratios, in spite of its high input expenditure.
3) Highest input expenditure but maximum of search effectiveness and survival power, optimum for intranet projects

Figure 2. Several strategies of information supply in comparison

Meaning disambiguation (# 1) is best assured in indexing of the controlled, best fitting, and integrated type. Automatic indexing fails here due to lack of effective interpretation, apart from specially programmed exceptions. In free indexing, the possibility of meaning disambiguation is provided but hardly used.

Predictability of essence selection (# 2) is only assured in the three aforementioned approaches of intellectual indexing, with the proviso that this indexing is based on a set of conceptual categories (see section 9). Otherwise, much uncertainty prevails concerning the essence selection and, correspondingly, concerning the choice search parameters. The consequence is low ratios of precision and recall.

Predictability of essence representation (# 3) is fairly well assured, but only in the variations of controlled, best fitting, and integrated indexing, with some reservations in case of merely controlled indexing. Predictability is reduced here. This is due to the freedom in term selection on the part of the indexer.

Paraphrase lexicalization (# 4) requires the knowledgeable interpretation of the texts to be made retrievable (see section 8). This is an indeterminate process also, that is, one that inherently defies any adequate algorithmization. except perhaps for a group of selected examples.

The possibility of *ellipses filling* (# 5) is also restricted to the three aforementioned interpretation-based modes of indexing.

Nontextual information (# 6) such as presented in images, graphs, and so forth, can reliably be made retrievable only through careful and knowledgeable interpretation – all unjustified claims to the contrary by some groups of artificial intelligence notwithstanding. This interpretation separates the immaterialities of an image from its essence and lexicalizes the aboutness of an image or any other graphical document (or document part) and expresses the essence in a predictable retrieval-useful manner.

It is true that free indexing provides the possibility of image coverage also, but since there is no index language in use here, the textual representation in the search file is unpredictable and therefore only slightly useful for retrieval.

Synonym control (# 7) is seldom encountered when no index language is in use. But sometimes it is algorithmically executed either during input or during search. Therefore, extractive and free indexing were assigned the reduced advantage indication "(+)."

Descriptor and keyword specificity (# 8) is highest in the integrated approach because here the reduced ratings in all the other strategies sum up to adequacy. Specificity is downgraded in case of full text strategy and in extractive indexing because the terms, highly specific as they may be, lack predictability. The same holds true for automatic indexing. In controlled and "best fitting indexing," the specificity is downgraded because the index language vocabulary is often not sufficiently specific here.

Recency (# 9) is greatest when any type of interpretation – which is necessarily an intellectual and time-consuming step – is omitted. Controlled and best fitting indexing perform worst here.

An adequate *possibility of retrieval-suitable syntax* (# 10) only exists when there is careful and knowledgeable indexing and also complete familiarity with the (necessarily artificial) index language grammar.

It is true that algorithmic text processing also makes use of syntactical-grammatical devices but, due to the uncertainty and unpredictability of their employment through the authors of texts, their usefulness for searches is limited.

The *possibility of collecting descriptor candidates* (# 11) requires the capability of managing uncontrolled natural language terms. This applies only where an index language is in use and the system also manages uncontrolled natural language terms. This is the case only in the integrated approach.

The *retrieval availability of terms that escape indexing* (# 12) applies only where there is no indexing at all and where indexing is complemented by the non-indexing strategies, as is exclusively the case in the integrated approach.

Input and search economics (# 13) are discussed in the conclusion.

14 Conclusion

We have tried to expound the difference between the goal of natural language on the one hand and that of an index language on the other, and to show the capabilities and limitations of the technology of processing natural language. On the basis of this analysis we show in which specific manner both these types of language can complement each other and for which specific purpose each of them is either appropriate or inadequate.

In launching an information system or in evaluating the suitability of a system for the purpose at hand it is important to have in mind which of the requirements enumerated in Figure 2 deserves priority or can be neglected. The entire architecture of an information system depends on such an analysis.

Here, the question of the costs incurred in the development and maintenance of an information system deserves particular attention. Natural language technology is often works than the knowledgeable and careful human. But the question is whether the goal of a sufficiently precise and complete information supply can be attained in an exclusively mechanized approached. If text interpretation is involved, mechanized text processing is bound to be unsatisfactory due to the inherent indeterminacy of this process.

The problem is also that an information system needs not only to meet the requirements that have been experienced and can easily be viewed in hindsight. Much more, an information system must also continue to serve its purpose in the more or less distant *future* and under conditions that may be dramatically different from those prevailing. The search files will have grown in size, as will have the frequency and specificity of the search requests to be executed and many new concepts will have had to be managed in storage and retrieval.

An information system which works to satisfaction in the experimental state may entirely fail when it is exposed to the conditions of every-day practice and when it has developed into large scale conditions ("small system syndrome"). The reason is that most of the experimental systems are memory-based, even if only latently.[26] They often make demands of the expenditure of time, patience and enduring attention on the part of the searcher, which cannot proportionately grow with the growth of an information system. Then it will have to be taken partly[27] or entirely out of commission. At that time, the entire enterprise reveals to have been a tremendous waste of time and manpower although in its beginning it seemed so economical and parsimonious in the purely empiricistic and positivistic view.

The consequences are dramatic. Not only is all the work lost that has been invested into the system in the past (or at least a steadily increasing part of it); in addition, in attepmting a new beginning one also faces the nightmare task both of coping with a stream of incoming new documents and also of re-doing what has already been done in the past. Furthermore, the loss of confidence on the part of management may well result in an insufficient supply of resources for a more promising new start up. Ironically, information system survival power has only very rarely been included in information system evaluations, when they have been executed in the presently dominating empiricistic- positivistic philosophy.[28]

An information system of high survival power excels when it takes into account the steadily increasing demands of the *far future,* and those, in fact, *already existing in the present.* Hence, at almost any point in time and in a less farsighted view, such an advanced information system will always appear overdeveloped and unnecessarily expensive in input, in particular from the accountant's viewpoint. A task not to be underestimated is convincing management that this seemingly exaggerated effort is a must for the enduring usefulness of the information system that must be financed.

Vendors' advertisements and researchers' success stories ignore, conceal or even deny the weaknesses of their text processing software and an inexperienced management is easily seduced into replacing human work through machine programs merely for cost reasons. It requires much entrepreneurial oversight and subject knowledge to find an adequate decision for the necessary (and early!) employment of sufficiently advanced conceptual and technical resources which are necessary to attain the goal of an information system of the desired *enduring* effectiveness and survival power.

Notes

1 Dahlberg is referring here to the philosophers Kant and Frege.
2 In our view, a concept is in existence *before* a natural-language word has been coined and even if no such term will ever be coined.
3 We are leaning here on v. Freytag-Loringhoff. p. 27.
4 cf. for example, Coates, 1960, pp. 19, 21.
5 "Known item searches" in present-day terminology.
6 "... that there are a myriad of ways of expressing even the most ordinary semantic content."
7 Axiom of predictability: "The accuracy of any directed search for relevant texts depends on the predictability of the modes of expression for concepts and statements in the search file."
8 "Indexing can be consistently wrong" (Soergel, 1994, p. 594). A patient is badly served through two doctors who *consistently* make the same but wrong diagnosis.
9 In our context, any language is regarded as "index language" in which there is *no free choice of expressions,* syntactical devices included. The terms may be of the natural language type (i.e., from a the-

saurus, optimally in specified meaning) or of the notational type from a classification).

10 The essence of a text must be separated from immaterialities, which for good reasons and according to the proved practice of decades of indexing, need not and should not be made accessible for retrieval.

11 "...indexable concepts had to be inferred from the text."

12 "Literary writing in particular is elliptical, purposely imprecise if not vague and full of implications and connotations."

13 "A document does not necessarily contain explicit information about its own subject."

14 International Classification 8 (1981), p. 96.

15 "Cutter's Rule" requires the employment of those expressions from an index language (in Cutter's time from a classification) which most appropriately and specifically represents the concept of interest.

16 Any advanced mode of indexing should be based on a set of conceptual categories. In a sevenfold manner they can constitute the inner skeleton of the indexing process (cf. Fugmann 1993, pp. 18-20; Fugmann, 2000, p. 19). A conceptual category is an extremely general concept above which there is no still more general concept in the field under consideration, for example, substance, process, and living entity.

17 "Axiom of fidelity": The accuracy of any directed search for relevant texts depends on the fidelity with which concepts and statements are expressed in the search file.

18 This holds true at least for the core field of an information system for which an index language has been designed.

19 This refutes the "inverse recall–precision" relationship, often cited as an allegedly inherent regularity in information retrieval. Such an inverse relationship is encountered only when indexing is executed in neglect of Cutter's rule or when any employment of an index language is dispensed with. Here the phrasing of query (or several of them for the same concept) constitutes an endless adventure of trial and error.

20 i.e., on a list of those words that must be extracted from the texts to be made better accessible for search parameters.

21 i.e., on a list of those words which are excluded from being filed.

22 This statement contradicts the presently widespread assumption that literally anything will be mechanized some day.

23 "The keywords approach with statistical techniques has reached its theoretical limit and further attempts for improvement are a waste of time."

24 An example is an attempt at locating documents on pythons being dangerous to man. Internet searches were executed in Alta Vista, Yahoo, Lycos, and Northern Light. A search for "python" produced between 110,982 and 412,410 responses. A search for pythons + suffocation + man with Boolean operators yielded between 22 and 8,687 answers. Most of the items with Boolean search were not relevant. The question of how many relevant documents escaped retrieval was not investigated.

25 It is even disputable whether such an type text processing deserves the designation of "indexing" because it does not lead to an index but merely to a concordance, i.e., to a list of text word occurrences. Extractive indexing does not meet the criteria for typical indexing as mentioned in section 9 and in International Classification 8 (1981), p. 96.

26 It is an indication of memory employment if one has to resort to the remembrance of author *names* in case of a search for a *topic* of interest.

27 "We only need the most recent literature."

28 This attitude has particularly emphatically been criticized by Budd (1990).

References

Bates, M. J. (1998). Indexing and Access for Digital and the Internet: Human, Database, and Domain. *Journal of the American Society for Information Science, 49*(13), 1185-1202.

Bernier, C. L. (1960). Correlative Indexes VI: Serendipity, Suggestiveness, and Display. *American Documentation, 11*, 277- 278.

Blair, D. C. (2002). Some thoughts on the reported results of TREC. *Information Processing and Management, 38*, 445-451.

Bloomfield, M. (2001). Indexing -Neglected and Poorly Understood. *Cataloging & Classification Quarterly, 33*(1), 63-65.

Budd, J. M. (1995). An Epistemological Foundation for Library and Information Science. *The Library Quarterly, 65*, 295-318.

Coates, E.J. (1960). Subject Catalogues -Headings and Structure. London: The Library Association.

Dahlberg, I. (1976). Ueber Gegenstaende, Begriffe und Benennungen (On referents, concepts and designations). *Muttersprache* Nr. 2.

FID/Classification Research. (1981). *International Classification*, *8*, 96.

Freytag-Loringhoff, von: Logik -Ihr System und ihr Verhaltnis zur Logistik. (Logics and its relation to logistics.) (4th ed.). Stuttgart: Kohlhammer Verlag.

Fugmann, R. (1985). The Five-Axiom Theory of Indexing and Information Supply. *Journal of the American Society for Information Science*, *36*(2), 116-129.

Fugmann, R. (1993). *Subject Analysis and Indexing – Theoretical Foundation and Practical Advice.* Frankfurt, Germany: Ergon Verlag.

Fugmann, R. (2000). Obstacles to Progress in Mechanized Subject Access and the Necessity of a Paradigm Change. In: W. J. Wheeler (Ed.), *Saving the User's Time through Subject Access Information* (pp.7-45). Chicago, IL: University of Illinois.

Gesellschaft fuer Klassifikation (Society for Classification). (1985). Free Text in Information Systems: Capabilities and Limitations. *International Classification*, *12*, 95-98.

Gopinath, M.A. see Ranganathan (1967).

Green, R. (1991). The expression of syntagmatic relationships in indexing: Are frame-based index languages the solution? In N. J. Williamson & M. Hudon (Eds.), Classification Research for Knowledge Representation and Organization. *Proceedinqs of the 5th InternationalStudy Conference on Classification Research* (pp. 79-88). Toronto, Canada.

Hjorland, B. (1997). *Information through Organization*. Westport, CT: Greenwood Press.

Ilmholtz, C. (1999). Review of O'Connor: Explorations in Indexing and Abstracting: Pointing, Virtue, and Power. *Key Words*, *7*, 10-12.

Milstead, J. see Mulvany, N. (1994).

Mulvany, N., & Milstead, J. (1994). Indexicon, The Only Fully Automatic Indexer: A Review. *Key Words*, *1*, 17-23.

Ranganathan, S.R. (1962). *Elements of Library Classification*. London: ASIA Publishing House.

Ranganathan, S.R., & Gopinath, M.A. (1967). *Prolegomena to Library Classification*. London: ASIA Publishing House.

Rijsbergen van, C.J., & Sembok, T. (1990). SILOL: A Simple Logico-Linguistic Document Retrieval System. *Information Processing & Management*, *26*(1), 111.

Roberts, R. (1997). Searching the New Dictionary of National Biography on CD-ROM. *SIDELIGHTS,* 11-12.

Soergel, D. (1994). Indexing and Retrieval Performance: The Logical Evidence. *Journal of the American Society for Information Science*, *45*(8), 489-599.

Shpackov, A. A. (1992). The Nature and the Boundaries of Information Science(s). *Journal of the American Society for Information Science*, *43*, 678-680.

Swanson, D. R. (1988). Historical Note: Information Retrieval and the Future of an Illusion. *Journal of the American Society for Information Science*, *39*, 92-98.

Weisgerber, D. W. (1997). Chemical Abstracts Service Chemical Registry System: History, Scope, and Impacts. *Journal of the American Society for Information Science*, *48*(4), 349-460.

Wellisch, H. H. (1992). The art of indexing and some fallacies of its automation. *Logos*, 369-376.