Robert Fugmann
Hoechst AG, Frankfurt/Main

The Complementarity of Natural and Indexing Languages

Fugmann, R.: The **complementarity of natural and indexing languages.** In: Int. Classif. 9 (1982) No. 3, p. 140–144, 33 refs.

It is a prerequisite of any successful literature search that one must be able to reconstruct or *predict* which modes of expression have been used in the search file to express the concepts or statements of interest. It is these expressions which must be looked up in an index or phrased as search parameters for mechanized retrieval.

With regard to *general concepts* the natural-language modes of expression, as used by the authors of documents, lack this predictability. It is inherent in any controlled indexing language or classification that it establishes representational predictability and, hence, prevents serious loss of relevant information, which would otherwise occur in retrieval. Sufficiently high retrieval precision can be attained through correspondingly large representational fidelity of the indexing language or classification. This requires well-balanced cooperation between vocabulary and grammar in these languages.

It is typical of *individual concepts*, on the other hand, that they are represented with good predictability and perfect fidelity even in the natural language of the author. Therefore, their translation into an indexing language is often superfluous.

Ambiguous "author-lingual" modes of expression should be preserved in the search file, too, although they may have already been represented there through an attempted indexing-language term.

(Author)

1. Introduction

Through the continuing expansion of on-line data bases many information seekers have become occasional information system users. Being primarily experts in their subject fields, they do not normally have the time or inclination to learn and persistently to memorize formal indexing languages for their literature searches. Therefore, those information systems which provide access to unrestricted natural language have attained considerable popularity.

User friendliness is, however, only one of several criteria which influence the acceptance of an information system. Effectiveness and accuracy, i.e. the capability of retrieving relevant information precisely and completely, is another and often decisive criterion. However user-friendly a system may be, it will have to be rejected if it does not meet the accuracy criterion in the long run.

The capabilities and limitations of indexing languages on the one hand and of unrestricted natural language on the other have often been investigated and commented upon (1) – (23), (32). But very contradictory opinions have been uttered on this issue, and we are still confronted with a most confusing picture. Several advocates of natural language information systems have gone so far as to ask, what could be more simple and effective than merely entering the wording of original texts into the data base without any revision and translation into an indexing language. Any inquiry should be satisfiable by such a file, because nothing is omitted, added or misinterpreted, and all the information is preserved that the human himself requires and uses for his relevance judgement.

Although information practitioners are well aware of the untenability of this argument if it is expressed in such a general form, it is not easy to refute this argument on the spur of the moment or at least specify the limitations of its scope of validity.

Often the ambiguity of unrestricted natural language is used as a counterargument. But there are natural, expert languages which provide markedly unambiguous expressions in their vocabulary, for example in chemistry and biology. In spite of that they have proved unsuitable for performing certain searches. For example, information on a certain class of chemical compounds or a class of living beings cannot reliably be retrieved from a data base which merely contains the names of individual chemical compounds or species of living beings.

Nor does the size of the natural-language vocabulary constitute a convincing counterargument. The vocabulary of the names of authors, individual chemical compounds, institutions, etc. can grow to any size, and in spite of this one will encounter no serious difficulties if the searches are restricted to those for authors, individual chemical compounds etc. On the contrary, such a vocabulary even provides a degree of (desirable) specificity such as is hardly equalled by the vocabulary of an indexing language.

Nor does the expense of entering unrestricted natural language texts constitute a prohibitive obstacle any longer. This holds true in particular if one is satisfied with abstracts or any other kind of document surrogates written in unrestricted natural language.

We shall in the following specify a property of language for which we shall claim indispensability if a particular kind of concept is to be handled satisfactorily in an information system. We shall also realize that the task of indexing languages can be defined as one of providing just this peculiar property. I am speaking here of the *predictability* of the modes of expression for the concepts contained in the data base.

In the light of the postulate of representational predictability (6), (33) the capabilities and limitations of unrestricted natural languages on the one hand and of indexing languages on the other will become more apparent.

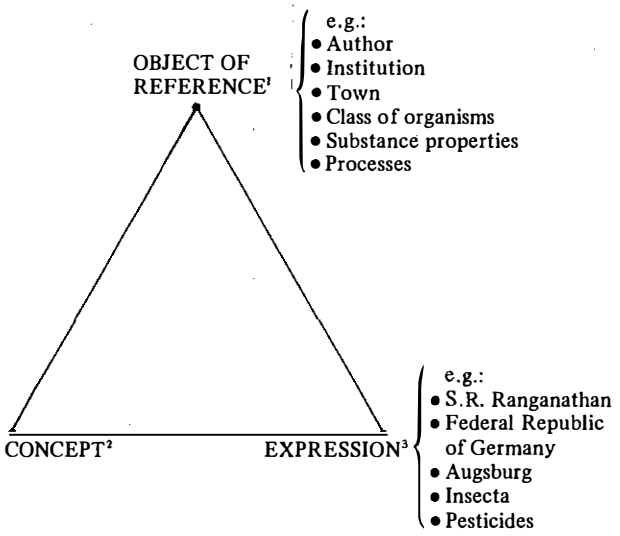
We shall base our considerations on a slightly modified version of the well-known Semantic Triangle of Ogden and Richards (26) – (30) as depicted in Fig. 1.

In the following discussion this triangle will serve to make clear what we consider to be the true goal of a literature search. Several controversies in our field have arisen merely because the authors were not aware that they disagreed on just this important factor.

2. The Goal of a Literature Search

The Semantic Triangle represents the relations between the "object of reference", the "concept" and its linguistic "expression".

A little reflection would reveal that of the three occupants of the triangle corners it is almost always the *concept* which is sought by an inquirer, and not the



1 Anything about which statements can be made.
2 The entirety of the true and essential statements about an object of reference.
3 Linear string of alphanumerical characters intended and agreed upon to convey a certain meaning.

Figure 1. The Semantic Triangle and the goal of a literature search

	CONCEPTS	EXPRESSIONS lexical	EXPRESSIONS non-lexical	
INDIVIDUAL	Author	S.R. Ranganathan	–	only
	Institution	Federal Republic of Germany West Germany	–	very few per
	Town	Augsburg	–	concept
GENERAL	Class of organisms	Insecta, (mosquitoes, beetles, ants . . .; 1 000 000 names	–	always very
	Substance property	Pesticides	. . . Substance X was effective against pests . . . . . . X suppresses the growth of weeds in wheat fields . . .	many per concept
	Process	–	Eradication of malaria transmitting insects by new pesticides	

Figure 2. Expression multiplicity for individual and general concepts

object or reference or the expression used to designate the concept or the object of reference.

"Concept" is here defined as the sum of the true and essential statements that are made or implied about an object of reference.

An "object of reference" is in this context anything about which statements can be made. Examples are an author, an institution, a town, a class of living beings, a substance property, a process of some kind etc.

3. Lexical versus Non-Lexical Expressions

It will also aid our analysis if we differentiate between "lexical" and "non-lexical" expressions for a concept under consideration (Fig. 2). A lexical expression is one which consists of a linear sequence of alphanumerical symbols, which by general agreement is used to represent a certain meaning. Examples are names of persons, institutions, geographical concepts etc. It is typical of lexical expressions that they always occupy one and only one place in an alphanumerical arrangement. Here, they can be logically entered and also later looked up.

Some consideration will reveal that a close relationship exists between the kind of concept and the mode of expression used to represent the concept. Thus, it is typical of *individual concepts* that they are expressed solely *by lexical expressions*, e.g. proper names. In this context, individual concepts are defined as being a kind of highly specific concept for which there is no more specific, still meaningful concept in existence, at least not in the subject field under consideration (cf (31)). For example, we do not know several species of an author, town, or institution. It is also typical that only very few of these lexical expressions are in common use for each individual concept, often only a single lexical expression. Therefore, for a concept in question these few lexical expressions can be readily looked up in dictionaries and compiled quickly and completely for an individual concept in case of demand. Retrieval is an important instance of such demand, which we shall soon discuss in some detail.

General concepts on the other hand are defined as being subdividable into more specific, still meaningful concepts. Thus, we know many different species of the general concept insecta, e.g. mosquitoes, bugs, butterflies, ants etc. What is of interest in the framework of our investigations is the *multiplicity of expressions* encountered for this general concept. There are almost a million names for the species by which the general concept insecta may be represented.

Another kind of expression-multiplicity is that prevailing for the general concept pesticides. In addition to this lexical expression an almost infinitely large number of non-lexical expressions is conceivable for this concept. Two of them are depicted in Fig. 2.

Still another kind of general concept is encountered exclusively in the non-lexical mode of expression. This holds true for the majority of all general concepts that occur in research work. An example is the eradication of malaria-spreading mosquitoes by a certain pesticide. A lexical expression for such a concept will be coined in natural language rather late, if at all. It is inherent in non-lexical expressions that they cannot be looked up in case of demand, because they defy effective alphanumerical arrangement. Any attempt to compile them will therefore constitute a long brain-racking search that must be terminated at some arbitrary point, long before this compilation is anywhere near complete.

But even if a general concept is represented exclusively by the names of the specific concepts which it comprises, the compilation of all these names is in most cases not feasible, owing to their very large number. Chemical compound classes are another example in addition to classes of living beings.

All these considerations have a strong bearing on the retrieval of texts relevant to the topic of an inquirer.

4. Representational Predictability as a Retrieval Requirement

If a document relevant to a topic of interest is to be retrieved, the inquirer must know how the topic was expressed in the search file. He will have to take this expression into account by means of a correspondingly tailored search parameter. If he omits this search parameter the document cannot respond to the query and will therefore not be retrieved. If there are different expressions for the topic in the file, each one of them will have to be taken into account as an alternative search parameter. If even one of them is omitted, loss of relevant texts is bound to occur, namely loss of those texts in which the topic was expressed solely by one of the forgotten expressions.

Let us now consider a search file into which the expressions of unrestricted natural language have been entered without any control of revision. Only in the case of an individual concept will the searcher be able to compile all those expressions by which this concept might be expressed in the file and which will therefore have to be included as search alternatives in the query. As we have seen, only in the case of individual concepts is the number of necessary alternatives reasonably small, so that all of them can be reliably and completely compiled.

This is in sharp contrast to a general concept, for it can be represented by an infinitely large variety of lexical or non-lexical expressions in a document of interest. Each of these possibilities would have to be included in the query as an alternative search parameter, if the relevant documents in the file are to be retrieved. Compiling and processing an almost infinitely large number of search parameters is, however, obviously impracticable. Hence a query for such a file will necessarily be phrased incompletely. Correspondingly incomplete will therefore be the responses to such a defective query.

Even the most sophisticated computer program cannot compensate for this kind of query incompleteness. No computer can be expected to work better than it was instructed to do by the query (and by the instructions that had been laid down in the program).

It has occasionally been argued that there is no need whatever to phrase the entirety of all conceivable expressions as alternative search parameters but only a small subset of them, namely those that have in fact entered the file. However, nobody is able to reconstruct or to *predict* (at the time of phrasing a query) just which of the many possible expressions happen to have entered the file, so that one could concentrate on them. An inquirer could restrict his search parameters to these expressions only if he had an opportunity of persuing the texts of relevant documents in the file before phrasing the query. However, these texts are exactly the ones still to be retrieved by the proposed search, and their wording cannot be assumed to be known in advance. Hence, except under artificial experimental conditions, searches for general concepts in a file for unrestricted natural language expressions will necessarily be incomplete, because of *the unpredictability of the natural-language expression for a general concept.*

This situation can be improved only by improving representational predictability. Seen in this light, *any indexing language constitutes a tool for achieving improved representational predictability.* Any classification, thesaurus, authority list, controlled vocabulary etc. serves this purpose. If more than only moderate demands are made on the completeness of the search for general concepts, then such tools are indispensible.

This is not refuted by several evaluation experiments which claim to have proved the general equivalence or even superiority of unrestricted natural language in information systems. These experiments were often conducted under markedly artificial conditions, such as small collection size, repetitive perusal of the test documents during the experiments, short period of coverage, restriction to only few authors (which results in an unnaturally high terminological homogeneity in the test file), and even revelation to the searchers of the wording of relevant documents to be retrieved. Under these unnatural experimental conditions representational predictability was not seriously put to the test. The test persons merely needed *to remember* the expressions which they had encountered in the recent past or even only during the experiment. Unrestricted natural language could therefore not display its distinctive weakness with respect to the predictability criterion. Therefore, it has always shown up unnaturally well in these experiments, often to the surprise of those who conducted these experiments and in particular of those who had had contradictory practical experience. It only added to the inconclusiveness of these experiments when no distinction was made between individual concepts on the one hand and general concepts on the other. Furthermore, pertinence as a precision criterion has often been confused with relevance.

5. Translation Problems

Using an indexing language of some kind is equivalent to translating the texts to be stored into this language. Translation of a text into another language may encounter considerable difficulties and always requires knowledge of the subject field involved on the part of the interpreter. If he has no general view of the vocabulary of both languages and of the meaning of each lexical unit he will be unable to find the most appropriate expression in the target language. Translation into an indexing language does not constitute an exception to this rule. If, for example, we are told in an original text that "silicosis" was observed, then the indexer must look for a species of "lung disease" in his vocabulary. Chemicals used for combatting Peronosporaceae in vineyards must be classified as "fungicides". "Pot life time" must perhaps be represented as "chemical reaction speed".

It has often been argued that during the unrestricted input of natural language expressions, taken from the original text, this expensive intervention of the specialist is circumvented. This argument, however, always requires completion, and the consequences of this kind of input should be mentioned as well, namely the incompleteness (and sometimes also imprecision) of the responses in case of a search for general concepts. Only if one can tolerate these deficiencies can expert knowledge be dispensed with during indexing.

6. User friendliness

Another argument that cuts both ways is user friendliness. On the one hand it is appreciated by the informa-

tion system user, in particular by one who is only an occasional user, if he can access the system with terms from the language with which he is familiar. On the other hand, his appreciation may well vanish when some day he realizes that all along he has been paying a high price for this comfort, namely loss of relevant texts in the case of general inquiries. Another price is the expenditure required for the guessing of possible expressions and combinations of expressions to reduce this loss, and also the price of the persistent uneasiness about the inherent inadequacies of all these efforts on the part of those who feel responsible for accurate literature searches.

7. Indexing language specificity

Another criterion in our comparison of information languages is specificity. In natural language we can express any desirable degree of specificity through the combined use of vocabulary and grammar, in particular through the syntax of the grammar. This is in sharp contrast to indexing languages, which almost always lack this degree of specificity. They have often been criticized for this deficiency. We must, however, ask ourselves the question, to what extent the high specificity of unrestricted natural language text can in fact be exploited by correspondingly specific queries. Detailed analysis reveals a picture of considerable complexity.

It is conducive to greater clarity if we distinguish between the specificity which a language displays to an intelligent and knowledgeable reader on the one hand and on the other hand that kind of specificity which is in fact available for retrieval. Much confusion has arisen because it was often falsely assumed that the high specificity of natural language text could be fully exploited through correspondingly specific search parameters. An analysis in the light of representational predictability will reveal that only a portion of the specificity encountered in unrestricted natural language text is available for retrieval:

Much of the specificity of natural language text is expressed in a non-lexical fashion, e.g. through *syntactical devices* such as prepositions, word sequence, pronouns, the grammatical cases etc. We have already mentioned that the multiplicity of expressions of this kind is almost infinitely large. It is therefore impossible to predict which special mode of expression might have been used in a document of interest. Thus, one would hardly dare to specify as a search parameter that, for example the word "against" should precede the word "pests" in a natural language sentence, or that both words must co-occur in a common sentence. The danger of losing relevant texts would be too large. Rather, one will in most cases prefer to omit any syntactical natural-language search parameter. This renders the query a fairly unspecific one, although the relevant texts may display a high degree of specificity, and the reason for this discrepancy is the unpredictability of the non-lexical mode of expression in unrestricted natural language.

An indexing language, on the other hand, may well comprise syntactical devices such as roles and links, relation indicators, relators, operators, topological graphs etc. They may enable us at least to express the relation between an organism combatted and the chemical substance that exerts the control. If these syntactical devices are sufficiently well defined (which has, admittedly, not always been the case) and reliably employed during indexing, then they can unhesitatingly be used as fairly specific search parameters without the risk of serious loss of relevant documents. Hence, an indexing language query may well be more specific than a corresponding query phrased for a file in unrestricted natural language. The reason is that the specificity of the indexing language surrogate of a document can be fully exploited in the query. This is due to the significantly higher predictability of non-lexical modes of expression in indexing language.

Another situation prevails with respect to the specificity of the *vocabularies* of both kinds of language. Natural language is often very fast in coining specific terms such as "insecticides", "pesticides", which make possible correspondingly specific searches in a natural language file. To take another example, natural language very early coined the terms "wet spinning", "dry spinning", and "melt spinning", but several indexing languages still content themselves with a more general descriptor merely for spinning. One of the reasons is that an indexing language vocabulary is normally revised only at long intervals. At least during this time until the language catches up with practice one will have to content oneself with less specific queries in the indexing language and with correspondingly less precise searches.

On the other hand, indexing languages need no public acceptance of new terms. There is, for example, no obstacle to introducing descriptors for concepts that would have to be designated as "mosquitocides", "beetlecides" or "anticides" in natural language. We also know the word "aquaplaning" in natural language, but not "oil-planing", "rotten-leaf-planing", "wet-clay-planing". In case of demand, however, indexing language can soon provide descriptors for these concepts and thus make possible fairly specific searches.

Their specificity will exceed that of searches in unrestricted natural language files, because, there, only non-lexical expressions are available. Many examples are found in languages like UDC or in the International Patent Classification. In these cases, it is the indexing language which provides more specific descriptors and achieves more precise retrieval responses.

Hence, it is by no means inherent in indexing languages that they are generally inferior to unrestricted natural languages with respect to specificity. Where such inferiorlty has been observed it has often been due to the avoidable primitivity of the indexing language under investigation, e.g. to its syntax deficiencies. This seeming inferiority may also have been due to the unreliable employment of such an indexing language, a phenomenon which may be closely related to its language-primitivity but which we shall not investigate more closely in this paper.

Let us, as an example, imagine an indexing language which employs a well structured, fairly specific vocabulary and a powerful syntax. Searches performed with such an indexing language will exhibit a degree of specificity exceeding by far that attainable with corresponding natural language files. At the same time the possibility of losing relevant information contained in the file is practically excluded. For the field of chemical substances and chemical reactions such indexing languages have already become reality.

In many other subject fields such favourable results may be unattainable with indexing languages, and un-

restricted natural language will be superior. Text ambiguity is one of the main reasons. This is the last topic in our evaluative comparison of both kinds of language in an information system.

8. Ambiguity of original texts

Authors often use expressions the purport of which is unclear. This gives rise to a variety of different meanings of such terms. Often the meaning that an author had in mind when using such a term cannot be inferred with certainty from the context. Owing to the assumptions that will have to be made about the meaning of such an expression considerable distortions may occur during translation. Indexing language as a target language does not constitute an exception to this rule. Here, these distortions will lead to responses that will often be regarded as irrelevant by an inquirer, in spite of all the effort made by the indexers. If in these cases the natural language term from the original text was entered into the file (at least in addition to an attempted, plausible indexing language representation) then the indexer cannot be blamed for this failure. He will also have been saved the time and the expense of performing this arduous translation.

9. Conclusion

Let me summarize the analysis as follows: If an information system is expected to deal with both individual and general concepts with more than only a moderate degree of accuracy, then both kinds of language can complement one another very effectively. Either of them must be employed just where it is most effective and must be dispensed with where its performance is typically inadequate.

In such a combination of information languages input expenses will inevitably have to be borne which are incurred by the knowledgeable and reliable employment of an advanced indexing language. Nor can we dispense with the employment of fairly sophisticated computer programs, mainly for the handling of the indexing language syntax.

However, if the only purpose of an information system is to search for individual concepts, the lower input costs may tip the scales in favour of unrestricted natural language, at least if searches have to be performed only infrequently. Otherwise it may be worthwhile to restrict the expressions in the file even for individual concepts in order to save the time required for compiling their different names for the query.

We have viewed the problem of natural vs. indexing language from the angle of representational predictability, and the question has turned out to be not one of either/or but one of both/and. This conclusion has also been reached by other authors with different arguments. It is largely due to the avoidable primitivity of our present indexing languages and classifications that they were occasionally rejected for a project under discussion. Overcoming this primitivity constitutes a rewarding goal for continued classification research and development, and it is in fact a goal equally as promising as that of exploiting the capabilities of unrestricted natural language more effectively.

References:

[1] Ranganathan, S.R.: Prolegomena to Library Classification, ASIA Publishing House, London 1967, p. 329 (ChapterMB).

[2] Lancaster, F.W.: Vocabulary Control for Information Retrieval, Information Resources Press, Washington, DC 1972, especially p. 1, 139.

[3] Soergel, D.: Indexing Languages and Thesauri, Melville Publishing Company, Los Angeles, CA 1974.

[4] Vickery, B.C.: On Retrieval System Theory, Butterworths, London 1968, p. 65.

[5] Fugmann, R.: The Glamour and the Misery of the Thesaurus Approach, Int. Classification 1 (1974) 76, especially p. 78.

[6] Fugmann, R.: Toward a Theory of Information Supply and Indexing, Int. Classification 6 (1979) 3, especially p. 6, 14.

[7] Gebhardt, F., Stellmacher, I.: Design Criteria for Documentation Retrieval Languages, J. ASIS 29 (1978) 191.

[8] Svenonius, E.: Current Issues in Subject Control of Information, Libr. Q. 47 (1977) No. 3, p. 326.

[9] Svenonius, E.: Natural Language vs. Controlled Vocabulary, Proc. Forth Canadian Conference on Information Science, London, Ontario, May 1976.

[10] Wellisch, H.: The Cybernetics of Bibliographic Control, J. Am. Soc. Inf. Sci. 31 (1980) 41, especially p. 45.

[11] Wall, R.A.: Intelligent Indexing and Retrieval: A Man-Machine Partnership, Inf. Process. Mgt. 16 (1980) 73.

[12] Lehmann, H., Blaser, A.: Query Languages in Data Base Systems, IBM Deutschland GmbH, Tiergartenstraße 15, D-6900 Heidelberg.

[13] König, E.: Verbindliches vs. freies Indexieren, in: Numerische und Nichtnumerische Klassifikation zwischen Theorie und Praxis. Proceedings der 5. Fachtagung der Gesellschaft für Klassifikation, Hofgeismar 1981, p. 263, especially p. 266, INDEKS-Verlag Frankfurt/Main, ISBN 3-88672-009-8.

[14] Wenzel, F.: Lösung morhplogischer Probleme im Frei-Text-Retrieval durch Segmentierung, Nachr. Dok. 30 (1979) 212.

[15] Haendler, H.: Selektionsgerechte Indikation von Sachgebieten und Sachverhalten, Int. Classification 2 (1975) 25, especially p. 28.

[16] Cherniavsky, V.: On Algorithmic Natural Language Analysis and Understanding, Inf. Systems 3 (1978) 5.

[17] Henzler, R.G.: Free or Controlled Vocabularies, Int. Classification 5 (1978) 21.

[18] Wellisch, H.: A Flow Chart for Indexing with a Thesaurus, J. Am. Soc. Inf. Sci. 23 (1972) 185.

[19] O'Connor, J.: Data Retrieval by Text Searching, J. Chem. Inf. Sci. 17 (1977) 181.

[20] Rogalski, L.: On-Line Searching of the American Petroleum Institute's Databases, J. Chem. Inf. Comp. Sci. 18 (1978) 9.

[21] Sudarshan, B.: Development of Reference Retrieval System with Simultaneous Building – up of Thesaurus for Industrial Information Centres, Libr. Sci. 16 (1979) Paper E.

[22] Olsgaard, J.E., Evans, J.E.: Improving Keyword Indexing, J. Am. Soc. Inf. Sci. 32 (1981) 71.

[23] Rothmann, J.: Online Searching and Paperless Publication, J. Am. Soc. Inf. Sci. 32 (1981) 71.

[24] Fugmann, R.: The Theoretical Foundation of the IDC System, ASLIB Proc. 24 (1972) 123.

[25] Mills, J.: Progress in Documentation, J. Doc. 26 (1970) 120, especially p. 123.

[26] Ogden, C.K., Richards, I.A.: The Meaning of Meaning, A Harvest/HBJ Book, Harcourt Brace Jovanovich, London 1936, p. 11.

[27] Sechser, O.: Modi der Bedeutung von Einzelausdrücken in Retrievalsprachen, in: Klassifikation und Erkenntnis II. Proceedings der 3. Fachtagung der Gesellschaft für Klassifikation, Königstein 1979, p. 1, INDEKS-Verlag Frankfurt.

[28] Dahlberg, I.: A Referent-Oriented, Analytical Concept Theory for INTERCONCEPT, Int. Classification 5 (1978) 142, especially p. 144.

[29] Dahlberg, I.: Zur Theorie des Begriffs, Int. Classification 1 (1974) 12.

[30] Dahlberg, I.: Conceptual Definitions for INTERCONCEPT, Int. Classification 8 (1981) 16.

[31] Freytag-Löringhoff, Bruno von: Logik – Ihr System und Verhältnis zur Logistik, Urban Taschenbuch 16, Kohlhammer, Stuttgart 1955, p. 27.

[32] Fugmann, R.: Naturalversus Indexing Language in Chemical Documentation, Ang. Chem. Int. Ed. Engl. 21 (1982) p. 608.

[33] Fugmann, R.: Role of Theory in Chemical Information Systems, J. Chem. Inf. Comput. Sci. 22 (1982) p. 118.