Heinrich Kaiser
Institut für Spektrochemie, Dortmund

Language Problems

Kaiser, H.: **Language Problems.**
In: Intern. Classificat. 1 (1974) No. 2, p. 87–89

In a thesis-like manner a number of proposals regarding economy in the presentation of information are presented among which in the first place one pointing out the advantage of replacing natural language expressions in favor of codes (decimal numbers) for the representation of concepts in a very precise way. Different abstracts services should be established according to the kind and value of information to be conveyed (rapidity on the one hand and critical absorption on the other). Finally, organizational problems of natural language translations of important but unknown papers are discussed.

(I. C.)

1. *The large variety of natural languages is generally regarded to be one of the main difficulties in creating a world-wide information system.*
This is only partly true: A natural language is necessary only for the very process of thinking, for person to person conversation, and in some cases for conceptual reasoning (e. g. discussion of results). All these operations are dealing with "open systems". However, as soon as ideas, facts etc. have been inserted into the closed hierarchic system of one of the natural sciences, using well defined concepts and quantification, *no natural language is needed for further processing of the information* in question. On the contrary, for most procedures of that sort a natural language is a rather improper tool, its words are signals with high redundance and open to connotations and misinterpretations. (The 'signal to noise' ratio is too low!)

2. For the processes of *compression and evaluation* indexing, abstracting, compiling, re-arranging, classifying etc.) artifically languages, or codes must be developed, which are less redundant, unambiguous, independent of natural languages. As symbols for concepts, facts and relations only pure decimal numbers should be allowed, in order to ensure compatibility within a world system.

3. A code, used as an artificial concise language for a definite, limited field of science can well be made to measure and does not need to be part of *one* universal code, presupposed some basic rules für accessibility and compatibility are obeyed. Translation into open text is no problem: any computer can be instructed to print out a translation of the number into a full description of the term in any desired national language.

4. From the foregoing it will be realized why at present some of the large printed or computerized information systems are not satisfactory: *The "intellectual interfaces" (connecting steps) between the human mind on both sides and the mechanized part of the system are in the wrong place.* Words of a natural language, or relics thereof, such as alphanumeric codes, are used in parts of the process where numbers only and pure logic relations should be at work. This results in large indexes, or in keyword-thesauri, in too complex operations etc. (It is like using a rapid computer with too slow a tape-reader or a typewriter).

5. The main reason for this embarrassing situation is a historic one: the co-operation of the users, specialized experts, in developing a systematic code was missing. It must be admitted that breeding a systematic code with a condensed vocabulary requires much intellectual effort and much patience, but it is worthwhile: eventually avoiding much more work of a similar character; less at a time, maybe, but at the wrong moment and with less effect.

6. Regarding *keyword thesauri and vocabularies*, the following facts should be kept in mind: the number of different words, each containing n syllables, which can be formed from an unlimited stock of Z different "syllables" (or symbols) is Z^n.

Most natural languages have vocabularies in the range from 300 to 30,000 basic words (if compositions of words are not counted). Such vocabularies could be built up from a stock of less than 300 different syllables and no word would have more than 2 syllables (300^2 = 90,000). For 30,000 words with three syllables a stock of only 31 different syllables or symbols is required!

7. What syllables are in a natural language, are the basic concepts (represented by symbols, which may be numbers or "keywords") in an artificial language. It is this stock of basic concepts which is meant when the term "controlled or restricted vocabulary" is used. With only 100 symbols in groups of 5 it is possible to form 10^{10} different expressions, more than ever will be used for indexing and abstracting.

The inventory of expressions in an artificial language (code) built up from a restricted vocabulary of symbols is accessible because it is formed in a systematic way. Only the basic concepts, their symbols and some simple rules on how to combine them must be memorized.

8. The active vocabulary of an average scientist may contain several thousand words. He understands more but he does not use more. Therefore keyword lists with several thousands words and more are of limited value only, even if they are superseded by some kind of systematic order. This is a bitter conclusion regarding the effort and money put into printed or into computerized information systems based on the words of a natural language. The computer can handle this, but the human user does not have all the words present in his mind.

9. With respect to the process of retrieval carried out by man, it mut be kept in mind that simultaneous performance of two or more additive acts of cognition does not often – if ever – occur. There is some evidence of a "single channel cognition mechanism". *For this reason*

the human mind is very inefficient for multi-dimensional searching. This task should in future principally be assigned to mechanical, optical or electronic devices, which will solve the problem much better.

10. The common practice of scanning the current literature or abstract journals by eye in order to find the answer (or all answers) to a *definite* question is ineffective, uncertain and a waste of time. Means must be provided to make this completely unnecessary. (Such a statement does not apply to the pleasure of browsing through the literature; this may be useful, stimulating, providing background orientation).

11. One major reason for the language difficulties in science information lies with the *secondary services*. They, in particular the abstract journals, are suffering from a conflict of duties: people wish to have a rapid service for awareness which means a "retrieval abstract" – and at the same time an "informative critical abstract" reporting results and data in order to avoid recourse to the original publication. The latter only requires open text in a natural language. It should be realized that these two tasks are incompatible. Their solution requires different technical means and different types of organisation. An abstract service can either be rapid or critical and thorough, because the preparation of a critical abstract, partly replacing the original, requires much time for consideration comparison, translation etc.

The obvious conclusion is that the present type of abstract services must be split up in two: one rapid, with modern means for multi-dimensional searching with practically no text, the other critical, thorough, much more selective. These two services may be produced by the same Centre as a complement to each other. It is interesting to observe that this development has begun already in some cases; it should be encouraged by all means.

12. Making a "retrieval abstract" in short time requires a numerical code with a restricted, prescribed vocabulary especially made for the field of science in question. Such a code, being independent of natural languages, will enable the authors to assist in indexing their papers.

The indispensable general co-operation of active scientists with the secondary services will certainly face strong aversion in the beginning. Therefore this co-operation must be organized within the compass of small groups (specialized and probably national); it must be done with a minimum of administration and the work must be distributed in such a way, that the part attributed to an individual scientist appears to him largely as the byproduct of his daily critical reading. The controlled vocabularies mentioned above, and questionnaires will help to avoid the hard labour of formulating a text for an abstract and in that way also overcome language barriers.

13. The very common idea, that *full text availability* of all "important documents" at any place and any time in advance to any request is desirable, should be exorcised. This again leads to enormous multiplication of work. If by modern means of retrieval the source for a document has been placed, the means of telecommunication will allow to get the full text within a very short time, and also a translation into any desired national language – if necessary. In some cases, where extremely quick delivery is urgent, the costs may be considerable, but regarding the costs for a stock of *all* documents, this will be negligible.

(This is the place to point out, that the attitude towards an original text in the natural sciences is quite different from that in the humanities. In these the text itself is the object to be preserved; in science, ideas and facts are extracted from the original and preserved in compilations. The original text after having served its purpose is forgotten and only in rare cases it is preserved as a document of historic value. This is a fact which may be very shocking to librarians and documentalists who are educated to look at original documents with the respect and the pride of a collector).

14. The splitting up of secondary services into one part for retrieval and another part for detailed information about results, will greatly reduce language difficulties. The quantity of text to be written, translated, distributed will ble much less than with the practice of today. The compression factor of the present abstract journal has been estimated to be about 10 (ICSU-AB). The splitting may well result in another factor of 10, giving a total compression factor of 100. The simple reason is that retrieval by means of a code for multi-dimensional searching will be a first step towards selective withdrawal:

Papers with a poor content or of very general character will enter the multi-dimensional index only with some general concepts, and therefore will not be found if detailed information is required. (The preparation of an informative critical abstract may come later because in the meantime the rapid service will lead to the primary publication if this is required.

15. The second stage of abstracting should be of higher quality than the present procedure. It should be more in the direction of critical reviewing and such work may well give the material for an annual or bi-annual critical review. From this it follows that this work must be done by specialized experts who are able to see the crucial points of a paper. This again requires co-operation of the user in information processing.

This second stage will contribute also to selective withdrawal. If an original paper does not contain *enough new* material, ideas or facts, it will not get such an informative abstract. This procedure can be formalized and made independent of the fuzzy idea of scientific value. For instance, a survey paper given during a congress may be most valuable and interesting, but it will not be abstracted once more.

16. This system will be especially effective, if the same multidimensional index is used as for the "retrieval abstracts". In this case *no paper would be entirely lost*, but the fact that there is a critical abstract also, would be a positive statement about the importance of the original publication. This mechanism could be used to withdraw occasional papers, e. g. conference proceedings, internal reports from government or industry, which make up a great deal of the multiple publications, just by not giving them a detailed informative abstract.

The next step in this process of compression is the selection of papers to be reported in review articles.

It can happen that the quantity of informative abstracts and critical reviews may be small enough to allow for translation into several national languages on a reasonable economic basis.

17. Another procedure has been discussed which might help to overcome language barriers, namely the problems of how to draw attention to important scientific papers written in any language, and from anywhere in the world: In each country the "top learned society" in a special field or a National Academy of Science could make a proposal which publication in its realm it regards to be of high importance, *measured by world standards*. This proposal should go to an ICSU-Unesco standing committee, which would provide the translation into one or more languages, generally used in scientific communication. These selected papers should be printed in a special issue of a scientific journal having a wide circulation. Specialized journals of international character are to be preferred for that purpose.

The costs for these translations and the publication should be paid for by the international body, but the respective National Academy should be responsible for the selection. If this selection is not done carefully enough and should prove below world standard, this will lower the reputation of the responsible academy and of the author. Therefore, this mechanism will work as a rule. The number of papers proposed for this collection of outstanding publications could be allotted to the national bodies according to the percentage of publications from this country in the worlds scientific literature. By this procedure papers otherwise buried by language difficulties or in journals of low circulation could be made known world-wide. This might be an effective help for developing countries too.

Helmut Felber
Österreichisches Nomungsinstitut, Wien

An Outline of International Terminological Activities

Felber, H.: **An Outline of International Terminological Activities.**
In: Intern. Classificat. 1 (1974) No. 2, p. 89–91

Consise survey on the areas of work in terminology, existing in (1) establishing teminological and lexicographical principles (work done within the ISO/TC 37), (2) in preparation of terminologies in particular subject fields and (3) in documenting ongoing terminological work and results of such work. Regarding the latter the tasks of Infoterm are listed and the specific forms of terminological data documentation are shortly explained. (I. C.)

1. Definition

The term "terminology" has a double meaning, namely:
Terminology[1] is the aggregate of terms representing a system of concepts of a particular subject field
Terminology[2] is the theory of terminology, which is an interdisciplinary field comprising linguistics, logic, information sciences and individual subject fields.

2. Importance for information

Terminology is fundamental for information: on the one hand it serves to word the information, on the other hand terms are used for indexing, storing, and retrieving the information. Thus specialized vocabularies, called thesauri, are prepared as documentation languages.

3. Three areas of work in terminology

Terminological work can be divided into three areas, namely in:

3.1 Establishing terminological and lexicographical principles
3.2 Preparation of terminologies in particular subject fields
3.3 Terminological documentation

3.1 Terminological and lexicographical principles

This work is carried out by the Technical Committee 37 "Terminology (principles and co-ordination)" of the International Organization for Standardization (ISO). The Secretariat of ISO/TC 37 is held by the Austrian Standards Institute in Vienna.
So far, ISO/TC 37 prepared the following six ISO Recommendations and one ISO Standard: