DOCUMENT RESUME

ED 051 860   LI 002 920

TITLE   Proceedings of the International Conference on
                General Principles of Thesauri Building, Warsaw,
                23-27 March 1970.
INSTITUTION   Polish Academy of Sciences, Warsaw. Documentation
                and Scientific Information Centre.
PUB DATE   70
NOTE   191p.

EDRS PRICE   EDRS Price MF-$0.65 HC-$6.58
DESCRIPTORS   Conference Reports, *Indexing, Information
                Retrieval, Information Science, Information Storage,
                *Information Systems, Language Development,
                Languages, *Thesauri
IDENTIFIERS   Poland, Scientific and Technical Information

ABSTRACT
                The aim of the conference was to hold a working
discussion on problems essential to the building of thesauri, a
uniform comprehension of which would make it possible to build them
in a manner ensuring easier communication and exchange of information
between information systems covering different ranges of subject
matter, and using different languages. The conference discussed the
way essential terms such as "thesaurus," "descriptor," and
"ascriptor" would be handled. The fundamental elements of which
thesauri should consist were also discussed and special stress as
laid on methods of thesauri building, selecting and qualifying
descriptors, their structure, and interrelations and determination.
The conference discussed the building of monolingual thesauri and
some problems of polylingual thesauri and debated the Unesco
publication, "Guidelines for the Establishment and Development of
Monolingual Scientific and Technical Thesauri for Information
Retrieval: third draft." (Author/AB)

PROCEEDINGS

of the International Conference
on General Principles
of Thesauri Building

Warsaw, 23-27 March 1970

Documentation and Scientific Information
Centre of the Polish Academy of Sciences

WARSAW 1970

1

PREFACE

THE INTERNATIONAL Conference on General Principles of The-
sauri Building, organised in Warsaw from 23rd-27th March 1970
by the Documentation and Scientific Information Centre of the
Polish Academy of Sciences, was attended by 33 participants
from 13 countries. The aim of the conference was to hold a wor-
king discussion on problems essential to the building of the-
sauri, a uniform comprehension of which would make it possible
to build them in a manner ensuring easier communication   and
exchange of information between information systems covering
different ranges of subject matter, and using different langu-
ages.

This being so, the agenda provided only for a brief presan-
tation by the participants of their particular views on issues
which were in principle set out in the questionnaire, and most
of the time was spent in discussing these matters.   But other
questions could be and were raised.

With a view to clarifying the ideas and meanings ascribed
to particular terms, the conference discussed primarily the way
these terms are to be understood. The terms included the essen-
tial ones such as "thesaurus", "descriptor" and "ascriptor"
/"non-descriptor"/, and a great measure of unanimity was shown.
The fundamental elements of which thesauri should consist were
also discussed. Special stress was laid on methods of thasauri
building, selecting and qualifying descriptors, their structure,
interrelations and datermination.

The conference discusaad the building of monolingual the-
sauri and some problems of polylingual thesauri, debated the
UNESCO publication "Guidelines for the Establishment and Deve-
lopment of Monolingual Scientific and Tachnical Theaauri for In-
formation Retrieval: third draft", and tabled a number of amend-

2

ments and remarks. Varying points of view were expressed on
thesauri, their structure, methods of building and even their
very essence. As the causes of the differences were exposed,
there were chances to resolve them. Partial agreement on some
basic formulations was also reached.

In publishing the preliminary material on the conference,
its course and results, I would like to emphasise that these
were only made possible by the dedication, effort and goodwill
of the participants. I hereby thank them all. I would also
like to express my deep gratitude to Professor Janusz Groszkow-
ski, President of the Polish Academy of Sciences, for his active
support and for his opening address. And I must also praise the
important contribution to proceedings made by the conference
secretary, Mrs. Barbara Krygier.

Kazimierz Leski

LIST OF PARTICIPANTS

1. AITCHISON Thomas M.   Institution of Electrical   ENGLAND
                         Engineers, Savoy Place,
                         London W.C.2

2. BELLERT Irena   Zakład Lingwistyki Formalnej   POLAND
                         Uniwersytet Warszawski
                         Warszawa, ul. Oboźna 8

3. BOJAR Bożena   Zakład Lingwistyki Formalnej   POLAND
                         Uniwersytet Warszawski
                         Warszawa, ul. Oboźna 8

4. GABROVSKA Svobo-   Bulgarian Academy of Sciences   BULGARIA
   dozaria M.   Scientific Information Centre
                         at the Central Library
                         Sofia, "7Noemvri" 1

5. GÓRNA Grażyna   Federation Internationale de   HOLLAND
                         Documentation, Secretariat
                         General
                         La Haye, 7 Hofweg

6. GRAVESTEIJN Jacobus   Bureau de Recherches   FRANCE
                         Geologiques et Minieres
                         45-Orleans, BP 818

7. JANSEN Rolf   Badische Anilin und   FRG
                         Soda-Fabrik AG
                         67 Ludwigshafen,Ammonlabor,C 6

8. KRYGIER Barbara   Ośrodek Dokumentacji i Infor-   POLAND
                         macji Naukowej
                         Polskiej Akademii Nauk
                         Warszawa, Nowy Świat 72

9. LESKA Maria   Centralny Instytut Informacji   POLAND
                         Naukowo-Technicznej
                         i Ekonomicznej
                         Warszawa,Al.Niepodległości 188

10. LESKI Kazimierz   Ośrodek Dokumentacji i Infor-   POLAND
                         macji Naukowej
                         Polskiej Akademii Nauk
                         Warszawa, Nowy Świat 72

11. LLOYD Joel J. American Geological Institute USA
 2201 M Street, N.W.Washington
 D.C. 20037

12. MAIXNER Vítezslav Ústředi věděckých,technických CZECHOS-
 a ekonomickych infgrmaci LOVAKIA
 Praha 1, Konviktska 5

13. MALMSTEN Karl A. FID/SW CRG, Swedish CR-Group SWEDEN
 104 35 Stockholm 23,
 Box 23019

14. MANČEVA Stefanka S. CINTI BULGARIA
 Sofia, ul."7Noemvri" 1

15. MARLOT Lucien Centre Nationale de la FRANCE
 Recherche Scientifique
 15 Quai Anatole France,
 Paris 7e

16. MICHEJDA Zbigniew Uniwersytet Warszawski, IBIN POLAND
 Warszawa, Krakowskie Przed-
 mieście 26/28

17. MOJŽÍŠEK Josef Ústředí věděckých, technic- CZECHOS-
 kych a ekonomickych informaci LOVAKIA
 Praha 1, Konviktaká 5

18. MOLNÁR Imre Library of the Hungarian HUNGARY
 Academy of Sciences
 Budapest V, Roosevelt tér 9

19. POLETYŁO Mikołaj Centralny Instytut Informacji POLAND
 Naukowo-Technicznej
 i Ekonomicznej
 Warszawa, Al.Niepodległości
 188

20. ROBOWSKI Józef Instytut Celulozowo-Papier- POLAND
 niczy
 Łódź, ul. Gdańska 121

21. ROLLING Loll N. European Community LUXEMBURG
 29, rue Aldringer,Luxemburg

22. ROSENBAUM Zentralinstitut für Infor- GDR
 Hans-Dieter mation und Dokumentation
 108 Berlin,Unter den Linden 8

23. SCHMOLL Georg Zentrale Leitung für GDR
 gesellschaftswissenschaftliche
 Information und Dokumentation
 bei der Deutschen Akademie der
 Wissenschaften zu Berlin
 108 Berlin, Universitätsstr. 8

24.	SCHANCHE Grete A.	Studieselskapet for Norsk Industri Oslo 3, Forskningsveien 1	NORWAY
25.	SCHIFF Ervin	Hungarian Central Technical Library and Documentation Centre, Budapest VIII, Muzeum 17	HUNGARY
26.	SPANG-HANSSEN Henning	Danmarks Tekniske Bibliotek 1350 København K.Øster Voldgade 10	DENMARK
27.	TOMAN Jiří	Základní knihovna ČSAV Praha 1, Národni tř. 3	CZECHOS-LOVAKIA
28.	VARGA Dénes	Computing Centre of the Hungarian Academy of Sciences Budapest 1, Uri utca 53	HUNGARY
29.	WEEKS David C.	George Washington University - BSCP 2000 P St.N.W. Washington, D.C. 20026	USA
30.	WISNER Roswitha	Ökonomisches Forschunsinsti-tut der Staatl. Plankommission, Abt. Information und Dokumentation 108 Berlin,Unter den Linden 69-73	GDR
31.	WOJTASIEWICZ Olgierd A.	Zakład Lingwistyki Formalnej Uniwersytet Warszawski Warszawa, ul. Oboźna 8	POLAND
32.	WYSOCKI Adam	UNESCO-Division of Scientific Documentation and Information 7,Place de Fontenoy, Paris 7e	FRANCE
33.	WÓJCIK Tadeusz	Uniwersytet Warszawski Warszawa, ul. Krakowskie Przedmieście 26/28	POLAND

S e s s i o n   c h a i r m e n

23rd March; Afternoon Session	-	Jiři Toman
24th March; Morning Session	-	Loll N. Rolling
Afternoon Session	-	Joel J. Lloyd

25th March; Morning Session - Jacobus Gravesteijn
               Afternoon Session - Thomas M. Aitchison

26th March; Afternoon Session - Henning Spang-Hanssen

27th March; Morning Session - David C. Weeks
               Final Session - Kazimierz Leski

CONFERENCE ORDER

1. The role of thesauri
2. Terminological problems
   a. The definition of a thesaurus
   b. The definition of a descriptor
   c. The definition and name of forbidden terms
   d. What should be demanded in order to accept a term as a
      descriptor?
   e. The terminology of cross-references
3. Constructional problems
   a. Which elements should be included in a thesaurus?
   b. How should a descriptor be built?
4. Methodological problems
   a. Methods of building thesauri
     - for given thematic ranges
     - for overlapping fields /with other thematics/
     - of multilingual thesauri
   b. The methods of building descriptors
5. Conditions which descriptors and thesauri must fulfil in order
   to ensure their inter-branch and inter-language correlation
6. Conditions which must be fulfilled by descriptors and thesauri
   as tools for further development of information
7. Organisational problems
   a. How to organize the popularisation of methods agreed, to
      give them the level of recommendations
   b. How to organize work in order to facilitate or make poss-
      ible the building of correlated thesauri

OPENING ADDRESS

I take great pleasure in opening this conference, one of
the first devoted to the building and development of thesauri.

I am also very pleased to greet the representatives of coun-
tries with varying attitudes to the problems of science and
scientific information.

Despite all differences, the main tendency is towards co-
operation, especially in the scientific field.

Scientific information is, of course, the basis of research.
Without accurate information on the state of a given field of
science, without sufficient knowledge of trends in other fields,
without the recording of new achievements, scientific research
cannot be practically efficient -- it cannot yield good results.

Developments which should be known to scientific workers
and to economists take place in different countries, are worked
out in differing languages and in different fields. The disse-
mination of such information precisely and quickly, and meeting
all practical needs, is not a simple problem. Only systems
which have similar means of determining the content of documents
and an unambiguous method of inquiry can assure success.

And here I come to the main aim of today's conference. At
the present stage of development of scientific information and
technical methods at the disposal of the retrieval system, the
most effective methodological tool to enable the precise assess-
ment of information, documents and retrieval seems to be pre-
cisely the thesauri.

There is still a danger, however, that the thesauri elabo-
rated in various countries and fields spontaneously, with only
a small degree of co-ordinated activity, though streamlining
the information system within the frameworks for which they
are designed, may create additional linguistic, inter-branch
and even inter-institutional barriers against a free flow of
information.

The task before the members of this conference is to con-
tribute to overcoming such barriers, or at least to reduce them

to manageable proportions. This is a difficult and ambitious task. I believe, however, that considering even the practical difficulties which are yet to arise, a great deal can be done.

I also hope that the presence at this conference of representatives of UNESCO, UCSU and FID — organisations with great international authority — will contribute to overcoming those difficulties.

I am happy to have the opportunity to wish the debates every success — in terms of concrete results — and I also hope that even if the planned aim is not fully achieved, at least a worthwhile approach will have been made.

Now, since work is not the only thing in life, I also trust that you will spend your free time enjoyably, and may your memories of this visit bring you to Warsaw again in the future.

Once again wishing you fruitful discussions, I now give the floor to the organizers, and declare the conference open.

Prof. dr J. Groszkowski
President
of the Polish
Academy of Sciences

ANSWERS TO QUESTIONNAIRE ON THESAURUS PROBLEMS

Thomas M. Aitchison[x]

1. D e f i n i t i o n   o f   a   t h e s a u r u s

A thesaurus is an alphabetical listing of concepts /i.e.
descriptors / which provides structural and relational infor-
mation about the concepts. A list of terms which does not
include structural and relational information is not a the-
saurus, even though it includes details of synonymous terms.
It is merely an alphabetical list of subject headings or an
alphabetical list of descriptors.
- Which structural elements /semantic, syntactic, etc./
  should be included in order to be able to call a given
  construction a thesaurus?
  A thesaurus should include the following: -
  a/ Concepts /i.e. descriptors/ arranged in alphabetical order
  b/ Details about each concept, i.e.
     /i/ Synonyms, alternative word forms, near synonyms, etc.
          For example:-
          AUTOMOBILES
             UF Motor cars
          with reciprocal entry
          MOTOR CARS use
          AUTOMOBILES
          /UF represents "Use for"/
     /ii/ Hierarchical /or structural/ relationships.
          For example:-
          AUTOMOBILES
             BT Motor vehicles

_____
[x] Institution of Electrical Engineers, London.

                    NT Estate cars
              /BT represents "Broader term".
              NT represents "Narrower term"/
    /iii/ Relationships other than hierarchical. There are
           a wide variety of related terms, which may be
           roughly classified into groups, such as "thing/
           part", "thing/property", "process/agent", "thing/
           application" to name only a few.
           For example:-
             Thing/part
                DIESEL ENGINE
                RT Pistons
             Thing/property
                Pistons
                RT Wear resistance
             Property/process
                WEAR RESISTANCE
                RT Testing
           /RT represents "Related term"/

-   Which elements, factors, influence the organisation of
    the thesaurus?

    a/ Subject field
       /i/   General vs. specific subject fields.A thesaurus
             in a specific subject field will cover the sub-
             ject field in greater detail than the general
             thesaurus. Some relationships between concepts
             are peculiar to a specific subject field. The
             same terms in a general thesaurus might not dis-
             play the specific relationships.
       /ii/ Differences between specific subject fields.
             Some subject areas have more precisely definable
             descriptors than   others /i.e. "hard" versus
             "soft" language types/. In some subject areas
             synonyms abound, whilst in others they are not
             so common. In some subject fields the   related
             terms and generic structure are more obvious
             than in others.

b/ Economic considerations

If costs must be kept low, this influences the speci-
ficity of the concepts selected and the number of re-
lated terms which are introduced.

- How should the degree of complexity and the number of in-
formation items, contained in a thesaurus be evaluated?

The more specific the concepts selected and the more
highly "pre-coordinated" the descriptors, the more comp-
lex the thesaurus will become and the more terms it will
include. A thesaurus in which the terms are at a "low
level of pre-coordination" /with complex terms being con-
structed from these simple concepts at the indexing stage/
will have many fewer terms than those with "highly pre-
coordinated" concepts. The disadvantages of low pre-
coordination level concepts is that frequently-sought
concepts are not included in the thesaurus and the des-
cription of the subject field is incomplete. If a con-
cept is not listed, related concepts cannot be shown.
There is need for research to find the optimum level of
specificity /i.e. pre-coordination level/.

2. Is the concept of a thesaurus
sufficiently complete and
univocally· and exhaustively
defined, or does it
necessitate further analysis?

Further analysis is required on the following points:-
a/ Rules for the selection of related terms /RT/. At pre-
sent the choice of related terms seems to be made
haphazardly. Where thesauri are constructed on a classi-
fied basis some help is given in this problem, but there
are relationships which cut across hierarchical groups
which can only be identified by those with a good know-
ledge of the subject field. It may be that it is impos-
sible to draw up detailed rules for selection of related
terms.

b/ Selection of specificity and pre-coordination level of con-
cepts. /see above/

c/ Thesaurus /classification systems. Advantages/disadvantages
of combining thesauri with classification systems. Problems
in construction and use.

Advantages.

/i/ Thesaurus and classification in a combined system com-
plement each other. The classification provides a vi-
sual display of subject fields showing hierarchical
and other relations whilst the theaaurus acts as an
index to the classification and at the same time ensu-
res control or synonyms and indicates related terms
which cut across the hierarchies in the classified
schedules.

:/ A thesaurus/classification system is a multipurpose
tool which can be used for the arrangement of books on
shelves /because it includes notation/ and for conven-
tional classified catalogues. Using the descriptors it
may be used for post-coordinate systems and for compu-
ter searching.

3. What is the role of a thesaurus?

- direct use in information and retrieval systems

a/ For controlled retrieval languages. Provides list of
controlled descriptors for indexing and searching. It
also provides generic and relational information which
allows for generic posting at the indexing stage, and
manipulation of the question at the searching stage
/i.e. the search may be made broader or more specific
by use of hierarchical relationships shown in the the-
saurus - and more exhaustive by drawing upon related
terms/.

b/ Natural language systems. Thesauri can be used to
suggest alternative terms etc. in compiling search for-
mulations in natural language. A different type of the-
saurus may be required fo: free-text searching; but
no research has been done, as yet on this problem.

13

in development of scientific information.

The listing of scientific terms showing structural and re-
lational information reveels interdisciplinary reletion-
ships. Unfortunately thesauri are always lagging behind
scientific developments.

4. Methode of construction of thesauri

- **Methods of compiling thesauri**
  There are obviously many: this is s personal view.
  e/ Definition of subject field.
  b/ Classification of subject field into main groups or fa-
     cets.
  c/ Collection of concepts /descriptors/ in each subject
     group or facet. Terms obtained from the literature,from
     other thesauri and classification systems, dictionaries
     etc. - also from subject experts.
  d/ Tabulate data on each concept. Find synonyms, etc.
  e/ Arrange concepts in detailed classification within each
     main subject group: this will assist in distinguishing
     hierarchical and other reletionships.
  f/ Arrange terms alphabetically. Insert use entries and
     check that BT/NT and RT entries are reciprocal.

- **Possibilities and advantages of automation in compilation
  of thesauri**
  The advantage of automatic compilation is that it ensures
  full reciprocity and avoids errors. It also takes much of
  the manual drudgery out of thesaurus compilation. Automa-
  tion should also facilitate up-dating and allow frequent
  new editions.

5. Conditions which descriptors and thesauri must fulfill in order to ensure their irter-branch and inter-language correlation

Inter-branch oorrelation can best be achieved if the the-
saurus is oompiled with the assistance of a classification sys-

tem. This will reveal that many concepts have applications in
more than one subject area and may fit into several hierar-
chies.

6. Conditions which must be fulfilled by descriptors and thesauri as tools for further development of information

a/ Thesauri must be easily up-dated. They must be hospitable
to new material. Technology is changing so quickly that
thesauri can never keep up. There should be frequent new
editions. This should be less expensive with automated com-
pilation.

b/ The change to natural language retrieval systems /which is
likely to come about with increased mechanisation/ may de-
mand a new type of thesaurus. More synonym dictionaries
may be required, and at the same time a larger number of
specialised thesauri - which must be rapidly up-dated.

SOME OBSERVATIONS ON THESAURUS PROBLEMS

Rolf Jansen[x]

1. Definitions

Concept: Mental idea of material or immaterial object based on common characteristics which are usually formed by abstraction and found identical.

Term: Name given to a concept and consisting of one or more words.

Descriptor: Univocal representative of a concept in a documentation system. The descriptor can be a fixed term /"preferred term"/ or any other designation.

Thesaurus: For purposes of information storage and retrieval a thesaurus is an orderly compilation of concepts

o represented by as many synonymous terms as possible in one or more languages,
o in which homonymous terms are specially marked,
o in which a descriptor univocally represents a concept, and
o in which semantic relationships between concepts are registered.

Relationships can be derived from the definitions of the concepts.

2. Structural elements of a thesaurus

Before determining the organization of a thesaurus and its structural elements, it will be useful to consider the principal

role of a thesaurus in documentation. The main object   of a
thesaurus is to facilitate information retrieval.Therefore the
following truism holds:

A documentation aearch asks for concepts rather than words.
The questioner haa an idea of the facts wanted, but the way in
which the facts are expresaed in a document or file does not
usually matter much. Any arbitrary formulations in the   docu-
ment or inquiry must therefore be eliminated, i.e. reduced to
the level of the concept.

Concepts as elements of information retrieval are therefore
the baaic elements of a thesaurus. Because concepts are expres-
sed by words and a number of aynonymous terms may stand for a
single concept, the conceptual level must be specially identi-
fied. Other measures that enable terms to be clearly assigned
to concepts include:

o   Homonymous terms should be identified. Short additions in-
     dicating the different meanings will do away with homonymi-
     ty and establish clear assignments.

o   For retrieval purposes it is convenient to use a designa-
     tion that univocally represents a concept and is called a
     "descriptor". It is of minor importance whether this des-
     criptor is a preferred term, a concept number or a syste-
     matic notation, although of course a notation is convenient
     in that it indicates not only the identity of a concept,
     but also its hierarchic relationships with other concepts.

The basic information contained in a thesaurus   includes
not only means of identifying conceptual levels, synonyms
and homonyms, their assignment to concepts,and the designations
of descriptors, but alao the aemantic relationships between
concepts, especially the hierarchical relationships. They form
the essential framework in any organisation of concepts and play
a decisive part in information retrieval. Every question   is
bound to incorporate hierarchical elements, either because nar-
rower concepta have to be taken into account or because   the
inquiry is relatively broad, i.e. starts at a higher level to
avoid loase .

As regaids the construction of the hierarchy it should be
noted that conceptual reality is polyhierarchic in nature and

manifests itself in netlike connections. Polyhierarchy means
that a concept can be assigned not only narrower concepts but
also any number of brosder concepts.

Other useful information in a thesaurus includes foreign-
language equivalents, different spellings, definitions of con-
cepts or explanations, important sources of conceptuel informa-
tion and aspects of subordination of concepts.

An example of a thesaurus structure is described below. It
has been prepared at BASF and adopted by IDC Internationale Do-
kumentations-gesellschaft für Chemie[2]. For each concept all
pertinent information is registered together in a specific
sequence. To enable the information to be processed by computer,
every item starts with a symbol identifying the various entries.
The total amount of information for each concept is called a
"concept set". The various entries in the concept sets of the
IDC Thesaurus are listed in Table 1.

The concept entries were made as versatile as possible to
enable all details available to be readily incorporated without
loss of information and to provide for any future extension to
new fields of activity. The first entries are intended for the
registration of synonymous terms in different languages. At the
end of each concept set are the directly related concepts, always
represented by one of the synonymous terms. Polyhierarchical re-
lationships exist where a concept is assigned more than one
broader concept.

The middle of any concept set is reserved for additional in-
formation, such as definitions and sources. Under entry "K"
each concept is assigned to one or more concept fields which are
based on aspects of concept categories and enable concepts to be
preordered by subject groups as the thesaurus grows. Assignment
to concept fields is by means of predetermined symbols consis-
ting of two capital letters /e.g. EO = properties, optical/.The
IDC Thesaurus covers about 40 concept fields which are important
in chemical documentation. Figure 1 shows some concept sets ta-
ken from the concept area "Optical Properties".

Begriffsbenennungen	–	Concept Terms

B Benennungen in Deutsch: Synonyme und Quasi-Synonyme
   Terms in German: Synonyms and near-synonyms

C Benennungen in Deutsch: Unterschiedliche Wortformen und
   Schreibweisen – Terms in German: Different Forms of
   the Same Word and Different Spellings

E	Benennungen in Englisch	–	Terms in English	
F		Französisch	–	French
I		Italienisch	–	Italian
J		Spanisch	–	Spanish
N		Niederländisch	–	Dutch
P		Portugiesisch	–	Portuguese

Zusätzliche Information	–	Additional Information

D	Begriffsdefinition	– Definition, "Scope Note"
H	Hinweise zur Benutzung	– Instructions for Usage
K	Begriffsgebiets-Einteilung /2-stellige Symbole/
	Concept Field /Two-digit Symbols/
Q	Quellen-Kurzangaben	– Coded References

Beziehungsbegriffe	–	Relation Concepts
Übergeordnete Begriffe:	–	Generic Concepts:

O	Oberbegriff	–	Broader Concept
S	Verbandsbegriff	–	Total Concept
X	Bezugsbegriff	–	Reference Concept

	Untergeordnete Begriffe:	–	Specific Concepts:
U	Unterbegriff	–	Narrower Concept
T	Teilbegriff	–	Part Concept
Z	Zugehörigkeitsbegriff	–	Accompanying Concept
G	Gegenbegriff	–	Opposite Concept
V	Verwandter Begriff	–	Related Concept

/Begriffssatzende	–	Concept Set Ending/

Tab. 1 Angabenkategorien der Begriffssätze des IDC-Thesaurus
Table 1 Entries of "Concept Sets" in IDC Thesaurus

3. Computer methods in thesaurus compilation

Computer methods of processing thesaurus information will be described below with reference to the IDC thesaurus.

The concept sets are punched on cards and fed to the computer, which assigns a serial number /"concept number"/ to each concept set and carries out various checking operations such as whether

/1/ any terms fed in are already available in the thesaurus, or

/2/ any terms specified as relation concepts have themselves been defined as concepts, i.e. occur under the categories for synonymous terms in another concept set. If not, the programme adds the missing concept together with the appropriate reference and a concept number. Otherwise the computer checks to see whether the reciprocal relationship is available and adds it, if necessary /"mutual completion of concept sets"/.

Concept fed in		Check	
E	fuel	E	fuel oil
U	fuel oil	O	fuel

1st question →
2nd question →

The checking programmes ensure that all semantic relationships are completed and all terms, including those in the categories of concept relationships, occur as synonyms of a concept.

The completed information can be printed out in ordered arrangement, classified by concept groups and completed with an alphabetical list of the synonymous terms specified in the subject section. Fig. 1 is a specimen of this section, while Fig.2 illustrates the alphabetical register. The computer printouts save the trouble of setting up and maintaining a card index which in the initial phase would have to be continuously improved and completed.

A disadvantage of thesaurus organization is that the relationships given for each concept always indicate only one hierarchical level. Therefore a computer programme has been deve-

ANGAEENKATEGORIE K / EC   SEITE   4


```
001506 8   OPTISCHE AKTIVITAET
       B   OPTISCHE DREHUNG
       B   OPTISCHE ROTATION
       B   OPTISCH AKTIV
       E   OPTICAL ACTIVITY
       E   OPTICAL ROTATORY POWER
       C   VERMOEGEN, DIE SCHWINGUNGSEBENE LINEAR POLARISIERTEN LICHTES
           OPTISCH ZU CFEHEN
       K   EO
       Q   ABC 2.992 - BEI 347 - CZS 28, 404 - FI 39, 75 - GM 47.19.A -
           HW 3/2.429 - RI 36:16, 41, 213, 219 - ROE 5:3604 - SB 1A -
           STU2 345 - UL 13.42
       X   STEREGCHEMIE                                             000372
       U   SPEZIFISCHE DREHUNG                                      000436
       U   MOLEKULARES DREHVERMOEGEN                                001593
       U   LINKSDREHEND                                             001594
       U   RECHTSDREHEND                                            001595
       Z   MUTAROTATION                                             001596
       Z   ROTATIONSDISPERSION                                      001597
       V   MOLEKUEL-ASYMMETRIE                                      001525
       V   CIRCULARDICHROISMUS                                      001599
       V   OPTISCHE ISOMERIE                                        002169


001557 8   FARBAENDERUNG
       K   EC
       U   THERMOCHROMIE                                            000425
       U   PHOTOCHROMIE                                             000428
       U   PIEZOCHROMIE                                             G00422


001562 B   PHOTOLUMINESZENZ
       E   PHOTCLUMINESCENCE
       C   AUSSENDUNG VON STRAHLUNG ANGEREGT DURCH ABSORBIERTES LICHT
       K   EO
       Q   ABC 2.820 - EHS 8:841 - GM 47.31.A - ML 729 - ROE 6:2.3780
       O   LUMINESZENZ                                              000430
       U   FLUORESZENZ                                              000095
       U   PHOSPHORESZENZ                                           000431


001563 P   CHEMILUMINESZENZ
       B   CHEMOLUMINESZENZ
       C   CHEMILUMINESCENZ
       E   CHEMILUMINESCENCE
       D   AUSSENDUNG VON STRAHLUNG WEIT UNTERHALB DER GLUEHTEMPERATUR
           INFOLGE CHEMISCHER UMSETZUNGEN
       K   EO
       K   EC
       C   ABC 2.820 -  AC 65/572 -  CZS 224 -  EHS 8:840 - GM 47.31.M
           -  ROE 6:2.3780 -  UJ 16:328
```


Fig. 1   IDC Thesaurus, arrangement by concept fields, within
         fields by ascending concept numbers

THESAURUS-ALPHABET-REGISTER,BAND-NR. C9194   03.11.69

RT C01699   B   PHOTOCPLCRIERUNG

EO 003424 . B   PHOTOCPRCHIE

EO C0042B   E   PHOTOCPRCHISR

EO 00042B   C   PHOTOCPRCHISPUS

RT OC2341   B   PHOTODIMERISIERUNG

SF COC221   E   PHOTOGRAPHIC DEVELOPER

SF C01203   E   PHOTOGRAPHIC EPLLSION

ZZ COC22D   B   PHCTOGRAPHIE

SF 001203   B   PHOTOGRAPHISCHE EMULSION

SF 000221   C   PHOTOGRAPHISCHE ENTWICKLER

SF C00221   B   PHOTOGRAPHISCHER ENTWICKLER

ZZ 001C62   B   PHOTOGRAPHISCHER FILM

RB C01701   B   PHOTOICNISATION

RT 001701   B   PHOTOICRISATION

RE C01701   B   PHOTOIONISIERUNG

RT 001701   B   PHOTOICNISIERUNG

RT 001698   B   PHOTOISCMERISATION

RT 001698   B   PHOTOISCPERISIERUNG

UR 000520   B   PHOTOKCLCRIMETRIE

EG C01562   E   PHOTOLLMINESCENCE

EO 001562   B   PHOTOLUMINESZENZ

RB 001694   B   PHOTOLYSE

RE 001147   C   PHOTOLYSE IN FLUESSIGPHASE

RE 001694   E   PHOTOLYSIS

UP CC0519   B   PHOTOMETRIE

UR C00519   E   PHOTOMETRY

RT C01700   B   PHOTONITROSIERUNG

RT C01799   B   PHOTOPCLYMERISATION

ZN 001799   B   PHOTOPOLYMERISATION

RE C0C567   B   PHOTOREAKTION

Fig. 2 Alphabetical index of IC Thesaurus

23

```
GENERIC CHART OF THESAURUS NO. 0066          STATE 27-10-69        (G)    (UT2)    PAGE
          1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

000630 LUMINESZENZ / KALTES LEUCHTEN / LUMINESCENCE
                                    V 001579  LUMINESZENZANALYSE
                                    V 001580  LUMINESZENZ-SPEKTRALANALYSE

001561 U SENSIBILISIERTE LUMINESZENZ
001568 U PHOTOLUMINESZENZ / PHOTOLUMINESCENCE
000693 U FLUORESZENZ / FLUORESZENZSTRAHLUNG / FLUORESZENZFAEHIGKEIT / FLUORESZENZVERMOEGEN / FLUORESCENCE
                                    V 001599  OPTISCHER AUFHELLER
                                    V 001584  FLUORESZENZANALYSE
                                    V 001585  FLUORESZENZANALYSE
                                    V 001586  FLUORESZENZTITRATION
                                    V 001586  FLUORESZENZINDIKATOR
                                    V 001587  FLUORESZENZMIKROSKOPIE

000632 U SENSIBILISIERTE FLUORESZENZ
001582 U RESONANZFLUORESZENZ
001583 U STOKES-FLUORESZENZ
000491 U PHOSPHORESZENZ / PHOSPHORESZIEREND / PHOSPHORESCENCE
000633 U MOLEKUEL-PHOSPHORESZENZ
001588 U KRISTALL-PHOSPHORESZENZ
001589 Z MOLEKUELPHOSPHORE
001590 Z KRISTALLPHOSPHORE
001591 Z NACHLEUCHTDAUER
001563 U CHEMILUMINESZENZ / CHEMOLUMINESZENZ / CHEMILUMINESCENCE
000634 U SENSIBILISIERTE CHEMILUMINESZENZ
001592 Z CHEMILUMINESZENZ-INDIKATOR
001564 U BIOLUMINESZENZ
001565 U KRISTALLOLUMINESZENZ
```

Fig. 3 IDC Thesaurus, hierarchical arrangement of concepts

loped at BASF which on the basis of the relationship entries
enables the thesaurus concepts to be printed out in the usual
hierarchical arrangement. The individual concepts, represented
by concept numbers and synonymous terms, are listed systema-
tically in separate lines, hierarchical levels being indicated
by indentation. Capital letters preceding the terms indicate
the different types of subordination /cf. Fig. 3/.

4. Use of a thesaurus in a computerized documentation system

There is yet another aspect to the computer processing of the
IDC Thesaurus. The thesaurus is an integral part of a computeriz-
ed documentation system[2,3] for the storage and retrieval of con-
ceptual information. The alphabetically-sorted thesaurus tape is
used for computer classification of terms that have been fed in.
    A brief outline of the procedure is given below. The concep-
tual information which represents the contents of a document   is
written as clear text on coding sheets /"indexing"/.The wording
of the concepts is free within certain limits governed by rules.
The newly-fed terms are alphabetically sorted and then matched
against the alphabetic thesaurus tape. All new terms are printed
out and pass to the thesaurus specialist for checking, editing
and incorporation into the thesaurus. This is followed by compu-
ter checking, up-dating and arranging operations. By comparison
with the up-dated alphabetic thesaurus tape all terms on the file
tape can be replaced by concept numbers and later by hierarchical
notations and thus standardized.
    The procedure is illustrated in Fig. 4 in a simplified flow
chart which shows two cycle processes, the storage and thesaurus
cycles which meet where the two tapes are correlated. It will be
seen that the IDC Thesaurus is an open concept system which is
kept constantly up-to-date owing to the automatic feedback.All
improvements in and additions to the hierarchical structure can
be subsequently transferred to the entire file. In this way,not
only is the information kept up-to-date, but also the system can
be adapted to future requirements.
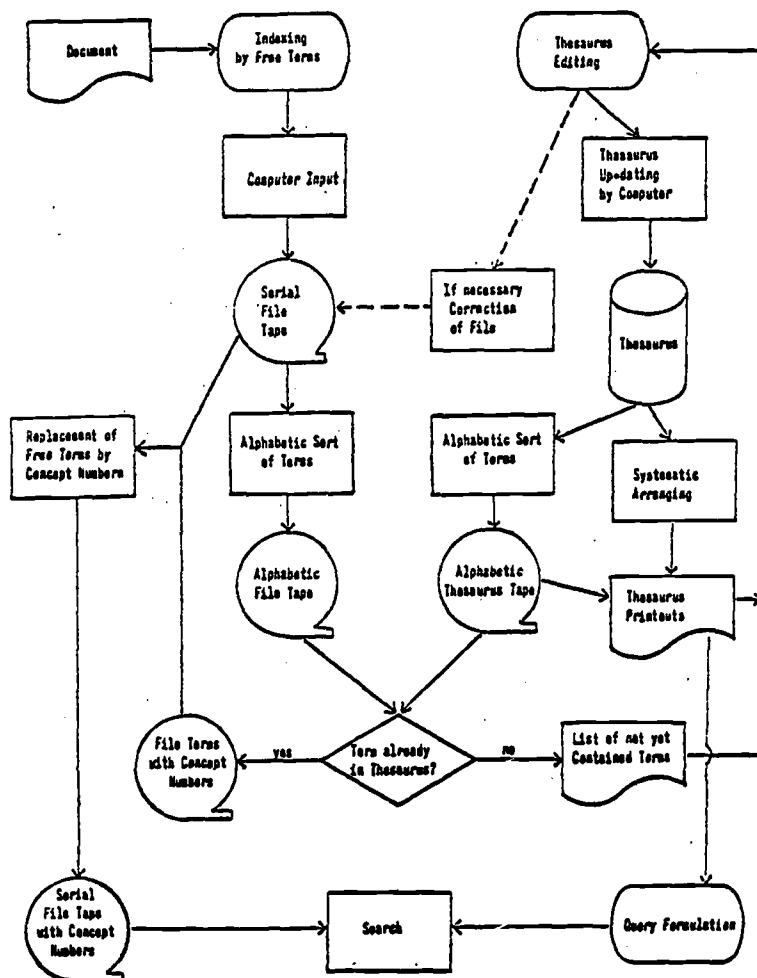
Figure 4. IDC Documentation of Concepts
Simplified Flow Chart

25

The fact that there is no fixed vocabulary is also an advantage. To cut out errors in formulating the contents of documents, particularly apt terms can be used and new concepts can be fed in without any delay. The thesaurus being multilingual, indexing is independent of language. Classification and coding of concepts have to be done only once, and then they take place automatically with always the same results.

5. General premises and problems of thesauri building

5.1 "Pre-coordinated" concepts or splitting into "post--coordinated" concepts

One of the more serious problems is the recognition and handling of compound or "pre-coordinated" concepts. By compound concept is meant a concept which can be mentally split up into separate concepts, the mental addition of which is certain to lead back to the initial concept. Accordingly, a single independent and unambigous concept is to be looked upon as the smallest conceptual unit whose further mental separation is pointless.

The problem apparently is that it is not possible to find an absolute standard for defining a demarcation between single concepts and compound concepts, the definition of demarcation depending on the user's point of view. In documentation the definition depends upon the requirements of information retrieval.

In a system of concepts the ratio between hierarchy and coordination is determined by the degree of subdivision of concepts. Concepts which are dissected do not exist in the thesaurus as such. Synonyms, definitions and hierarchical relationships cannot be registered. Dissection of concepts interferes with the hierarchy of the classification system.

Because dissection influences indexing and thesaurus building, it is necessary to establish certain rules taking into account the significance of hierarchy for the ordering of concepts and for information retrieval. On the other hand it is important to consider whether compound concepts, if any, should

be entered into both file and thesaurus if the combination of
their basic concepts /"post-coordinated" concepts/ can be stored
just as unambiguously and appropriately. In any case a clear
distinction should be made between the mental dissection of con-
cepts and the purely linguistic dissection of compound terms or
words.

Under the indexing rules of IDC the basic policy is to dis-
sect concepts only to the extent that the original conceptual
connection in the context is not lost and there is no change in
meaning. Each of the single concepts, when considered out of
context, should have professional information value and be de-
finite in scope.

5.2 Treatment of synonyms and homonyms

The question of conceptual unity also appears in connection
with synonymity, i.e. when it is to be decided whether terms
should be registered as synonyms or quasi-synonyms for a parti-
cular concept or whether they designate different concepts. In
linguistics the opinion is held that there are no true synonyms
and that each term stands for a new and independent concept. In
documentation, where the emphasis is on practical usefulness, such
rules are too strict. In checking terms for synonymity the que-
stion is not only whether differences in the definitions of the
concepts can be detected, but also whether usage differs in the
literature and whether any difference is relevant for informa-
tion retrieval purposes.

While synonyms are independent of context by definition, ho-
monyms depend on the context and should therefore be identified
as such during indexing. In the thesaurus the best way to cha-
racterise homonyms is to write a parenthetic expression behind
the term; this expression should contain a conceptual limitation
in the form of an explanatory word or at least a number. In the
IDC system the parenthetic expression causes the computer to
issue a message when the thesaurus is matched against the new
storage tape, so that homonyms may be recognized and brought into
an unambiguous form on the storage tape.

If thesauri of different scientific fields are to be compa-
tible or used in different branches of knowledge, an additional
overall check for homonymity is necessary.

5.3. Definition of concepts

As additional information on a concept, its definition is of great value. At least in cases where relationships are difficult to indicate, a definition should be worked out and entered into the thesaurus.

In the interest of practical documentation, concepts should be defined with general usage in mind rather than terminological standardization, however justified. Experience shows that most differences in the use of special technical concepts can be attributed to differences in the degree of abstraction of the underlying definitions. One author has a narrower, another a broader idea of the meaning of a concept. Because of the given scope of variation, it is of no use to fix unduly narrow definitions.

Taking all this into consideration, the following guidelines for a user-oriented definition of concepts result:

o The definition of a concept should be as close as possible to the usage of the concepts in the literature.

o In cases of doubt, the more abstract and comprehensive version is to be given preference in order to avoid loss of information during information retrieval.

5.4 Different kinds of concept relationships and their definitions

In order to describe concept relationships it is necessary to have an understanding of the different kinds of concept systems and the different types of relation concepts. Stimulated by a publication by WÜSTER[7] the terminology of the types of relation concepts has been completed and put to practical use in the IDC Thesaurus. A distinction is made between abstraction system /"Abstraktionssystem"/, whole-part-system /"Bestandssystem"/ and attributive or affiliation system /"Zugehörigkeits-system"/, each having two reciprocal types of relation concepts.

Abstraction systems are characterised by broader concepts and narrower concepts. Narrower concepts have all the characteristics of the broader concept and in addition at least one limiting characteristic.

Whole-part systems indicate the relation between an entity
and its parts and are described by total concepts and part con-
cepts. Part concepts are arrived at by the mental division of
a whole /total concept/ into its parts.

Example: "body", "chassis" and "engine" are part concepts
of the total concept "automobile".

Attributive or affiliation systems are described by refe-
rence concepts and accompanying concepts. Accompanying concepts
are closely related to their reference concepts, but dc not
coincide with them in their characteristics/no abstract relation-
ship/, nor can they be formed by a division or dissection of
their reference concepts /no whole-part relationship/.

Examples: "Catalyst" is an accompanying concept of "catalysis";
"Distilling column" is an accompanying concept of
"distillation".

These examples show that the attributive system includes im-
portant cross-references between different concept categories,as
between a process and the function of a material, or between
operations and apparatus, while in the abstraction system rela-
tionships are restricted to one and the same concept category.

In the system of concepts, abstract relations, whole-part
relations and attributive relations appear simultaneously and
penetrate each other. Therefore, if the order of the system is
to be readily understood and the organization is to be convenient
for retrieval purposes, it is important that the different types
of relation concepts be identified and arranged in a logical
sequence.

While in the abstraction and whole-part systems concept re-
lationships result directly from the definition of the relation
concepts, the direction of the relation in the attributive or
affiliation system is obtained by logical interpretation of
"affiliation". It is a characteristic of "affiliation" that it
depends on and belongs to something else; the meaning of "affi-
liation" is that it refers to an object which can be named as a
reference concept.

O	Broader Concepts	}	Abstraction System
U	Narrower Concepts		
S	Total Concepts	}	Whole-part System
T	Part Concepts		
X	Reference Concepts	}	Attributive System
Z	Accompanying Concepts		
G	Opposite Concepts		
V	Related Concepts		

Table 2   Kinds of relation concepts with abbreviation sym-
           bols of IDC Thesaurus

With compound concepts the situation is more complicated.
A general rule for the assignment of concept relations is de-
rived from the fact that compound concepts have been formed by
either "determination" or "conjunction" or "disjunction"   or
"integration" /cf. 1/. Most common is the formation by determi-
nation, i. e. limitation of the original concept by the addition
of a supplementary characteristic to the content of the original
concept. In this case the basic word in the compound word is the
broader concept. The explanatory portion is the reference con-
cept or, if the relation between the explanatory portion and the
compound word does not seem to be relevant for retrieval purpo-
ses, the related concept.

Example: E Oxidation catalyst
                          O Catalyst
                          X Oxidation

Therefore: E Catalyst   E Oxidation
           U Oxidation catalyst Z Oxidation catalyst

It is, however, not always easy to distinguish between
single concepts and compound concepts and if pre-coordinated,
the type of formation is not always apparent.
Concepts for which it is difficult to determine the direc-
tion of the relation are characterised as "related concepts".

These are concepts that cannot be assigned to any of the other
relation concepts and whose cross-relationship often consists
in a somewhat remote mental association.

An unresolved problem is how to deal with overlapping. In
the IDC Thesaurus overlapping is indicated by special referen-
ces.

6.   O u t l o o k

The only way to avert the impending disaster in information
handling is through international cooperation. What is required
above all is agreement on the methods to be used. Efficient and
future-oriented methods of documentation must be developed. It
is being realized more and more that the efficiency of a docu-
mentation system depends upon the perfection of its conceptual
system.

A thesaurus seems to be the best means to compile concepts
together with all pertinent information and to generate a poly-
hierarchical system of concepts. The consistent use and precise
representation of concept relations determines the quality of
the thesaurus and its reliability in information retrieval.The
problems of finding the right methods for constructing a the-
saurus and the many decisions which are connected with the study
of special concept fields have only been briefly dealt with in
this paper.

S e l e c t i o n   o f   r e f e r e n c e s

1/ DIN 2330: Begriffe und Benennungen. Allgemeine Grundsätze.
   Juli 1961

2/ E.Meyer und R.Jansen, in Vorbereitung

3/ E.Meyer: The IDC System for Chemical Documentation. J.Chem.
   Doc. 9 /1969/, S.109-1 3; Die IDC und ihr Dokumentationssy-
   stem. BASF-Information, Sonderausgabe Sept. 1969

4/ M.Scheele: Wissenschaftliche Dokumentation. Schlitz/Hessen:
   Eigenverlag 1967

5/ D.Soergel: Klassifikationssysteme und Thesauri.Frankfurt/
   Main: Deutsche Gesellschaft für Dokumentation 1969

6/ G.Wersig: Eine neue Definition von "Theseurus". Nachr. Dok.
   20 /1969/, S. 53-62

7/ E.Wüster: Die Struktur der sprachlichen Begriffswelt und
   ihre Darstellung in Wörterbüchern. Studium generale 12 /1959/,
   S. 615-627

REMARKS ON THE PROBLEM OF THESAURI AND THEIR BUILDING

Maria Leska[x]

Definitions

1. A descriptor is a chosen formalised term, consisting of one or a few words or a determined symbol, forming an elementary part of the thesaurus, meant to represent in a univocal way the determined subject content, of equivalent terms or groups of terms.

2. Ascriptor — a term or qualification largely synonymous with the given descriptor. Ascriptors in the information retrieval system are replaced by corresponding descriptors,while in thesauri they indicate corresponding descriptors by means of a system of reference marks.

3. A thesaurus is an orderly, arranged quantity of notions and termini creating an open system of subject headings placed within the framework of a domain or a problem, classified, part of which consists of descriptors with indicative interdependences and their mutual conceptional relations.

Structural elements of a thesaurus

1. General scheme of groups and subgroups of descriptors.

2. Index /may be tabular/ of descriptors taking into account their mutual hierarchical dependences, semantic and functional connections, ascriptors etc.

---

[x] Central Institute for Scientific, Technical and Economic Information, Warsaw.

3. An alphabetic index of all terms embraced by the thesaurus with cross-references showing the correlation of the terms, their mutual connections and the connections between descriptors and respective ascriptors.

4. Description of the system of cross-references applied in a thesaurus.

Note: A list of terms with neither structural information nor information about their mutual dependences is not a thesaurus, even if it has partly synonymous terms.

A thesaurus should as a rule cover fully the given domain or problem for which it was created. However, analysing existing thesauri and the literature concerning them, we can select two tendencies:

1 - a thesaurus should be adapted to a definite collection of information material and serve to make accessible the necessary information from the respective collection without either information "noise" or silence.

2 - a thesaurus should fully embrace the respective domain or problem, whether or not in the informational collection to which it is applied the materials are complete. Such a thesaurus would help in tracing the information gaps of respective collections.

The definition of a thesaurus is a problem which still demands discussion. It would be most advisable to define univocally the conception of a thesaurus, for which as we know there are numerous definitions. The importance of differences in the way of formulating may be here accepted as negligible; more important, even essential, are differences in the substantial understanding of this conception. A certain number of authors consider that a thesaurus is a kind of dictionary: ideological, or notional. In such cases the requirements given to the thesauri are limited to:

- obtaining a certain number of words of natural language chosen by analysis of the subject matter of texts and systematised according to an initially chosen classification system,

- obtaining an index of descriptors as a tool for a precise designation of content of the necessary information and enabling coordinate indexing of documents and informational inquiries,

- creating an orderly collection of key words chosen by statistical analysis of texts, no matter what subject.

The point of view does not sufficiently explain the still growing importance that is attached to the thesaurus, eepecially as to the information retrieval language in modern information systems and in studying modern methods of scientific information. There is a basic difference between a thesaurus and a dictionary. A dictionary is used to find words and terms, and a thesaurus - notions. We therefore cannot consider a list of terms which does not include structural information about mutual connections and dependences of notions as a thesaurus.

In the latest definitions a thesaurus is described also as a classification system. This view is interesting, but needs discussion. It is indispensable to designate what additional conditions should be accomplished by a thesaurus in order to be generally considered as a classification system independently of its main aim as an information retrieval language.

It seems that the reasons quoted above sufficiently justify the need of continuing further discussion.

Our experiences and the analytical and research work carried on while preparing a thesaurus of scientific information, show that certain classical methods are being formed: of choosing informational material for the thesaurus, of defining the scheme and its thematic range; principles of working out descriptors; the system of cross-references.

Each thesaurus still has special features which distinguish it from the others. This is due to the domain or problem for which the thesaurus has to be built, and to the concrete conditions in which it is worked out and applied.

The building of a thesaurus should be preceded by an accurate designation of the aim which it should serve and by the designation of available criteria of the quantity of notions, which were collected for building the thesaurus. The methodology of building the thesaurus depends also on the accepted definition of the thesaurus.

The first stage of work upon a thesaurus, the operation of collecting the optimal quantity of notions concerning the domain or problem is a classical stage, and there cannot be any possibility of avoiding it.

How varies the methods of collecting an assemblage of no-
tions may be depends, however, on the range of the thesaurus.
- Considering that a thesaurus is adapted only to a desig-na-
ted collection of informational material, the quantity of no-
tions may be formed by collecting key words and groups of words
systematically from the documents of the respective collection.
- Assuming that the thesaurus should embrace a given domain
or problem, it is necessary to construct - already as the first
stage of the work - a semantic-hierarchical scheme of this the-
matic range.
Contrary to the first solution, when solving the problem as
above, the operation of gathering notions and terms should not
be based on one information collection, even on the best one.
Such a collection and the materials embraced in it should be ana-
lysed systematically and very thoroughly. The results should be
compared with other local and foreign collections of materials,
with semantic-functional-hierarchic schemes of this domain or
problem, theoretical ones as well as those based on experience.
It is also necessary to investigate the largest possible number
of information items - and retrieval systems of this domain or
problem.
When the work on building the thesaurus in conducted paral-
lel to the process of gathering information material /in the
range of the domain or problem/, a third solution could be ap-
plied. It should be started with elaborating a theoretical
semantic-hierarchic scheme. This scheme will be more theoretic
than when building it on the basis of an existing and systema-
tized collection. Such a solution gives the possibility of ex-
cluding, already at the design stage, materials only loosely
connected with the basic thematics of the domain. And such mate-
rials are always to be found in every collection.
Gathering an amount of notions and termini should be reali-
zed on the basis of the natural language, embracing in the col-
lection entries specific to the specialized language of the do-
main or problem and even if necessary terms originating from slang.
When planning analytical and research work on the fund of
notions and terminological problems of the domain, the limits of
this domain should be precisely defined.
It is especially important that the thesaurus should be built
for the range of the domain, and not for the given collection.

The delineation of the scope of the domain for which the thesaurus should be built and the comparative analysis of existing semantic schemes of this domain and of related domains will be of evident aid when delineating the thematic range of the thesaurus.

The solve the problem of designing the optimal way of gaining access to the collected notions and termini, one has to classify the collection and to range the notions and termini according to the categories of the scheme. The goal of this operation is to find out the gaps in the fund of notions and termini and to correct and improve the scheme.

During this stage of the work it is already necessary to define the principles of utilising the collected mass of notions and termini in the thesaurus, and especially:

- the criteria of choosing termini to become descriptors,
- the principles of descriptor building,
- the principles of utilising synonyms, closely related words etc.,
- the principles of connecting descriptors with ascriptors.

These criteria and principles should enable the elaboration of a collection of descriptors and ascriptors showing their relationships. A tabular display with descriptors systematized according to the scheme of the domain or problem may be used here as an efficient tool.

The determination of the method of alphabetic ordering of descriptors and ascriptors is to be done in the next stage of the work. An arrangement in the form of a permutated index should be most effective as far as informative value is concerned. The system of cross-references should correspond to the hierarchic relationships of descriptors and their interconnections with ascriptors shown in the tabular display.

I do not intend here to discuss all essential problems arising when buildi: the thesaurus - only the most essential elements.

The popularity of thesauri as information retrieval tools is still rising. This is evidently a sign of their usefulness. Nevertheless the difficulties bound up with elaborating the-

sauri are great. Although the methods of building thesauri are
still more precise, it is impossible to build a thesaurus fol-
lowing precisely these methods: each thesaurus has its own
problems, and demands modification and complementation of the
method. In order to master all these difficulties it is neces-
sary to collect experiences. This may be done by means of pub-
lications devoted to the direct exchange of know-how and of
experiences concerning the theory of building thesauri as wall
as their practical application. So far, experiences on the ef-
ficient use of thesauri as ratrieval tools have been amassed on
the basis of a very limited number of thasauri. Such experien-
ces help to evaluate tha thesaurus in its role as:
  . - a vocabulary of notions enabling the correct indexing of
documents,
    - an information-retrieval tool,
    - a clasaification system.
    It seems also of importance to discuca how the development
of thesauri will influence the development of scientific infor-
mation theory. It is also to be dasired that the development of
the theory and principles of thesauri building will eventually
reach the stage of universalization and internationalization.

PRINCIPLES OF THESAURI BUILDING

K. Leski[x]

The concept of thesaurus that I propose is intended to em-
brace all thesauri, not only those built for domains with al-
ready highly-formalised languages /as in technology, the exact
sciences etc./.

x

1. **T h e   c o n c e p t   o f   a   t h e s a u r u s   a s
   a n   e l e m e n t   o f   a   r e t r i e v a l   s y s-
   t e m**

The goal of building thesauri is to obtain a univocal way
of qualifying the content of documents and of their retrieval.
The univocal character should be achieved in what concerns the-
matics and organisation on a possibly large scale. The range of
thesaurus thematics should therefore be as wide as possible,and
it should embrace all the collections of materials from this
domain.

The optimal thesaurus should therefore - theoretically -
comprise the whole of human knowledge and should be applied to
all retrieval systems.

At the present development stage this is not possible, for
several reasons. One could say that it is rather likely   that
such a stage will never be reached. While a necessary precision
and univocal character of the terminology, the identity or at
least the semantic and functional similarity of terms examined

---

from all points of view, will be achieved, the knowledge of
laws governing the living languages will probably increase so
that it will be possible to organize the retrieval systems on
different bases and with the aid of tools more efficient than
the present ones.

The conclusion concerning the range that respective the-
sauri have to cover is that they should comprise the ranges of
subjects where the respective elements of the thesaurus are
looked upon from one point of view.

Such ranges are practically represented by respective
fields of science, and thesauri should therefore be adapted to
their frameworks.

· There are no limits to the organizational reach of the the-
sauri. It is languages only which may form such barriers: the
relations between the notions in various languages that were
formed in different cultural conditions may be differentiated;
therefore translating notions from one language to another may
cause either great informational "noise" or silence.

Within the frames of one language or one domain such a
danger doesn't exist. We can assume in such conditions that a
thesaurus should comprise the range of a full domain of science
and in these frameworks, all the collections applying retrieval
systems based on methodological tools worked out in the given
language.

The thesaurus cannot present an inflexible construction -
stiff, closed and fossilized. It must be receptive to the inf-
low of new notions covering new problems, be amenable to the
removal of out-of-date terms, and be able to absorb changes of
meaning in the descriptors. It must therefore present an open
system.

· The thesaurus must ascertain an almost automatic passing
from terms embraced by a certain problem to other problems con-
nected semantically and functionally.

It must therefore reveal the relations between the terms
comprised /or between notions represented by these terms/.

As a result of this reasoning /obviously abbreviated here to
fundamental elements/ the thesaurus may be defined as follows:
"An open system covering a determined full thematic range con-
taining an ordered number of terms, some of which are admitted

as descriptors, showing the relations occuring between these terms and their mutual dependences".

Such a thesaurus consists of 2 categories of terms: descriptors and other terms. The thesaurus cannot exist without these terms.

Theoretically, we can assume that a thesaurus can embrace all the semantically-precised terms from a certain domain. Evidently such a multitude of terms would be too numerous and as a consequence such a thesaurus would cause a lot of information noise or silence, and above all its application would involve much difficulty. The thesaurus must therefore embrace only some such terms, and only those where the level of generality will be high enough.

Among these terms may appear some groups with semantically close meanings etc., therefore there won't be any necessity of applying all these terms. It will be advisable to choose terms representative for such groups. The method of building such terms cannot be quite free - they must be subordinated to some common rules.

While epplying these terms it is not possible to refer to the whole thesaurus each time in order to find their mutual relations. These relations, at least the most important, must be indicated directly by each of them. The most effective way of doing this is by means of cross-references.

All the terms cannot be applied operatively while they can be met within the given thematic range, and therefore their roles must be determined in the thesaurus. Thus we come to the definitions of fundamental elements of the thesaurus: descriptors and ascriptors.

A descriptor is a "chosen formalized term, built of one or e few words or a symbol, making an elementary part of the thesaurus intended to represent in a univocal way the determined subject content, with pointing out its basic relations and dependences with other descriptors and ascriptors".

An ascriptor is "every term included in the thesaurus, but not a descriptor, being one of the synonymous terms, or one close in meaning in relationship to one descriptor comprised by the thesaurus".

2. Structural elements of a thesaurus

The abovementioned way of comprehending the essence of the thesaurus and its elements creates special requirements for its construction. As a consequence, such a thesaurus should comprise:

a. A general scheme of groups and subgroups of descriptors.

b. A tabular display of descriptors, showing their mutual hierarchic dependences, their semantic and functional relations, ascriptors etc. [1]

c. An alphabetically permutated register for all terms applied in the thesaurus, indicating their mutual relations by means of cross-references.

Such a register may be divided into separate lists of descriptors and ascriptors.

Besides the above forms of presenting a thesaurus, the relations between the terms applied in it may be presented by other methods. However a system which does not comprise the forms mentioned above will not be a thesaurus.

3. Elements influencing the organization of the thesaurus

The abovementioned structural elements of the thesaurus determine its external shape. Its subject content, and as a consequence its organisation, depend on:

a. The thematic range comprised by the thesaurus and its semantic content.

---

[1] The tabular display of descriptors, realised according to the scheme comprising all dependences and relations of each descriptor /encl.1/, fulfills two roles simultaneously: of a material enabling a clear and full picture of these dependences and relations and of the analytica¹ apparatus, controlling the building of the scheme of groups and subgroups of descriptors. The subject division of the thesaurus can also be established in another way, but the already-mentioned analytic method appears however to have many positive aspects; above all it eliminates various interpretations.

b. The assumed degree of minuteness of the thesaurus,clos-
ely connected with the method of operating the thesaurus.

c. The overlapping of the thematic ranges in the thesaurus
with other thematic ranges.

d. The clearness and terminological precision of the the-
matics embraced by the thesaurus, and as a consequence the met-
hod of choice and building descriptors.

4. Building of the thesaurus

A thesaurus should be built so as to ascertain as full an
analysis as possible of its thematic range. Its thematic range
should be therefore determined and presented in the form of a
semantic-hierarchical scheme.

Such a scheme should then be filled with subject entries
collected on the basis of analysis of documents entering within
the range of the planned thematics, books and periodicals, pub-
lished and not published, of inquiries addressed to the collec-
tion and of different types of analyses of frequency and means
of appearing in the documents of those entries.

Filling the scheme should be followed by correcting and
improving the established semantic hierarchical scheme. A very
important stage in the building of a thesaurus is to work out
which ranges it has in common with other thematic ranges.These
will be the ranges through which will take place the flow of
information between the thematic range of the respective the-
saurus and neighbouring thematic ranges. While analysing the
subject entries appearing in these ranges, it is necessary to
look at them from two points of view /or more/ of the overlapp-
ing subject ranges. Similarly, we should approach the choice and
building of descriptors in two or more ways, establishing their
relationships etc. Great help can be provided here by an accura-
te building of tabular display of descriptors.

5. Language barrier

In item 1 I have mentioned the language barrier as limita-
ting the reach of organisational application of the thesaurus.
Such barriers undoubtedly exist and overcoming them will always
be difficult, but may not be impossible.

It is obvious that the semantic contents of respective sub-
ject entries, and as a result of descriptors, sometimes differ
from one language to another.

The semantic hierarchical arrangements of these entries and
descriptors, as well as of their mutual connections and correla-
tions, the synonyms and homonyms etc. are differentiated.

When passing from the system of a thesaurus built for one
language to such a system in another language, it is necessary
to consider carefully all these differences and create thesauri
common for these languages. There will evidently be no possibil-
ity of obtaining in this case a very high level of conformity
of descriptors and their arrangement. It is to be assumed that
the precision and minuteness of a multilingual thesaurus will
be somewhat limited.

The degree of conformity will here be the function of all
these factors and probably a result of their given interpola-
tion.

6. The role of the thesaurus

In item 1 I limited the role of the thesaurus to the role
of an element, a tool in the information-retrieval systems.

Such is probably the contemporary role of the thesaurus.But
is it the only role? Such a limitation would not be right. A
properly-built thesaurus is meant first of all to qualify do-
cuments, and later to retrieve then. Even at this first stage
another possibility of utilising a thesaurus reveals itself:
analysing the content of documents with the help of descriptors,
and in consequence breaking up these documents into microele-
ments, classifying these microelements and determining their
mutual correlations. The question arises of whether the the-
saurus we are building now will perform the role of an instru-
ment for such an analysis, or whether another type of thesaurus
will originate from the objective systems of documents in the
given domain.

7. The correlation of thesauri

According to the above reasoning, it seems that there is
the possibility of a free flow of information from one thematic

range embraced by one thesaurus to another thematic range.Such
a result may be achieved through:

   a. Agreement on understanding the essence of the thesaurus.

   b. The unification of requirements from a thesaurus,concern-
ing necessary structural elements for its existence and ways
of building thesauri and descriptors.

   c. Paying particular attention to overlapping thematic ran-
ges common to interrelated thematics and examining the descrip-
tors embraced by these ranges from all possible points of view.

   d. Making thorough comparisons of thesauri systems with the
same /or similar/ thematic ranges, but in different languages,
and building bi-lingual or multilingual thesauri.

   Satisfying the above conditions should ascertain the corr-
elative character of newly-built thesauri.

   In order to achieve a similar result in relation to existing
thesauri, it is necessary to analyse thoroughly their ranges
common with other domains and thessuri - if possible also adap-
ting descriptors created for these ranges.

   Though the new possibilities of using thessuri ss instru-
ments for analysing the contents of documents /item. 6./do not
seem to have a direct influence on their mutual correlation
this aspect of the problem deserves close attention nevertheless.

	The examined descriptor	1
	The definition-designation of the examined descriptor	2
	The descriptor of directly broader content	3
	The head descriptor of the group	4
	Descriptors subordinate directly to the examined descriptor	5
	Descriptors associated with the examined one as to subject and function	6
	Head descriptors of descriptors associated with the examined one	7
	Other domains of science with which the examined descriptor may be related	8
	Synonyms of the examined descriptor	9
	Homonyms of the examined descriptor	10
	Terms close to the examined descriptor	11
	Resulting ascriptor	12

Tabulary Display of Descriptors

Encl. 1

REMARKS ON THE GENERAL PRINCIPLES OF THESAURI BUILDING

Imre Molnár[x]

1. Hierarchical and alphabetical thesaurus

Ordered collections of the terms of human knowledge, which contain an enumeration of concepts, as well aa their interpretation and relations, are named thesauri. This definition includes a few elements which need further explanation.

1.1. Ordered collection of terms

As shown by the title, the appraisal of human knowledge can be realized by collecting the terms, on the one hand, and the systematization as a condition for a collection of terms designated as a thesaurus, on the other. A thesaurus may be constructed in several ways, using different principles of arrangement. A thesaurus compiled for a certain purpose owes its particular structure to the general relation between structure and function. Besides, the developmental stage of the thesaurus is also a factor determining the structure. A thesaurus haa differently structured forms in the course of its development. In general in might be said that the first and fundamental form of thesauri is the hierarchically-arranged structure. This form reflects the aubordination of science. The hierarchically-ordered thesaurus gives an interpretation of the correlation of terms in a particular branch of science, and makes possible an appraisal-in-depth relating to special questions of different scientific problems.

---

[x] Library of the Hungarian Academy of Sciences, Budapest

46

If a thesaurus has the function of gathering all possible concepts of a discipline, the hierarchically-ordered form is best suited to this task.

A thesaurus considered as an instrument of information systems can hardly be effective in the hierarchical arrangement. The hierarchically-ordered list of terms is not suitable for information analysis, indexing, storage or retrieval of information. These problems may only be solved by alphabetically-arranged thesauri. The hierarchical structure of a thesaurus compiled for information work is only the first step to the final, alphabetically-arranged form. The hierarchical thesurus remains, of course, a necessary instrument of a complete thesauri-system because only this type of thesaurus is able to give adequate answers to such questions as these:

Are the terms of a branch of science sufficiently detailed?

Are the markings of different correlations of terms suitable?

Does the thesaurus contain synonyms, homonyms, etc. in a necessary quantity?

Does the thesaurus contain a certain new term? If so, what kind of relations does that term already have?

Thesauri which are used in the routine work of information systems and are often born during continuous information work, are employed in alphabetical form. By its mechanical arrangement, as well as by its codes for relations of terms and by its homonyms separated and synonyms connected by code signs, the alphabetically-arranged thesaurus provides an efficient instrument for the indexing of documents, for information storage and retrieval, and also for standardization of information queries.

There are many possibilities of marking the hierarchical level and semantic relations of included terms in the alphabetical thesaurus. The codes connected with each term of the thesaurus are able to represent and mark the class of terms which the actual term belongs to.

This is shown by a detail of a biochemical thesaurus.

H i e r a r c h i c a l arrangement		A l p h a b e t i c a l arrangement	
12	Protein	Dehydrogenase	1211
121	Enzyme	Enzyme	121
1211	Dehydrogenaae	Lactate-dehydrogenase	12111
12111	Lactate-dehydrogenaae	Protein	12

Both the code number of every term and the number of digits of a code number have a definite meaning relating to the hierarchy of terms.

1.2. Human knowledge

Thematically, the main types of thesauri are: general and special thesauri. The type of a thesaurus is thematically defined by the eize of the branch of science which the concepts belong to. The limitation of different disciplines and of their terms seems to be one of the most exciting queations of thesauri building.

The completenese of the terms of a discipline is always relative, the limita of a branch of ecience cen be differently appraised in depth.

1.3. Interpretation of terms

Every term in a thesaurus must have its exactly-defined and characteristic meaning. Therefore, terms must be interpreted exactly. This interpretation may be different in aize: it may include only the meaning of the terms but it may also contain synonyms, related terms, etc. It is also advisable to interpret whether e term is a descriptor /indexing term/ or only a subject-heading /no indexing term/.

The criterion of unanimity requires, in the first place, the separation of homonyms. The use of brackets is a frequent method of making the necessary distinction between the different meanings of a homonym.

1.4. Marking of relations

Semantic correlations mark the hierarchical level of terms in a thesaurus. The relations lead from e generic term to a

specific one and vice versa. The most frequently used rela-
tions are as follows:

     Broader term /BT/

     Narrower term /NT/

     Related term /RT/

     Use   /USE/

     Used for   /UF/

  The sufficient marking of both the hierarchical level and
the relations of terms in a thesaurus ensures that the requ-
ired depth in the indexing process is achieved.

  2. T h e s a u r i   i n   i n f o r m a t i o n
         s y s t e m s

  Information systems and the indexing process frequently
result in a thesaurus. However they always have a significant
feed-back to the thesaurus. An information system can only
be effective if it corresponds with the continuous and rapidly-
changing information needs. This is why any one branch of sci-
ence needs a special thesaurus for the use of a certain instit-
ution, for the different depth of information queries. A re-
search problem of a discipline may be of more importance in
one particular place than in another. The dispersion of usage
of certain terms is the greatest in borderline sciences. An
institution which studies a problem regarded as an important
one analyses its terms in a more detailed way than another
which tends to investigate this problem only peripherically,
and uses its terms only in a more general break-down.

  This situation leads to the atomization of the methods and
techniques of thesaurus building. And this atomization has no
negative character in general. For all the institutes, research
laboratories, enterprises, etc. it is a duty and a necessity
to build their own thesauri which must correspond with the in-
formation needs in their own organization.

  This development offers fewer and fewer opportunities for
thesauri which tend to cover the entire range of human know-
ledge. The rate of the production and specialization of infor-
mation makes it more and more urgent to build thesauri, rela-
tively narrow in volume and analytically deep in appraisal.

These thesauri may be compiled and structured in a variety of
depths, sizes, forms, etc.

On the basis of the abovementioned arguments, all the
principles, facts, methods and techniques elaborated and made
usable for thesauri building are today very valuable and ne-
cessary. Preparation of data, methods, techniques and designs
is the greatest help experts of thesaurus research may give to
the specialists of information systems.

3. General thesaurus, special thesaurus

The construction of a thesaurus requires a comparatively
long time, and time is also the destiny of the thesaurus. As
a consequence of the acceleration of scientific development,
the division of the individual disciplines and the semantic
range of the terms will be transformed, will broaden out or
will become narrower, but in certain cases they may cease.

The division of the content of terms causes most problems,
often those most difficult to solve. A scientific problem
which was expressed during a certain period by a single des-
criptor may function in the future as the content of several
descriptors.

This process makes for a rich system of cross-references;
however this richness later becomes pullulation, which makes
the system at first cumbersome, and finally useless.

As a consequence of the abovementioned process, thesauri
must be rearranged periodically. A thesaurus can be expanded
during a relatively long interval; this possibility has,however,
a decreasing occurrence in time. The possibility of expansion
is inversely proportional to the growth of a thesaurus.

In general, thesauri have a retrospective character. The
larger the thesaurus is, the greater is the validity of this
statement. Computers may result in a significant time-saving
in the compilation of very large thesauri; there is, however,
much to do which is not mechanisable on computerizable, and
these tasks require long and skilled labour so that the the-
saurus will be retrospective anyway.

The larger the volume of a thesaurus, the harder the prob-
lem of its application to modern and analytical information
needs, owing to the inevitably poor representation of the
emerging new scientific fields of our days.

It is important to make efforts to build thesauri for the
individual branches of science. The building of relatively small
thesauri is always more economical than that of larger ones.
This is why I feel it necessary to interpret the basic princi-
ples of thesauri building according to the size of different
types of thesaurus.

НЕКОТОРЫЕ ПРОБЛЕМЫ КОНЦЕПЦИИ ТЕЗАУРУСА

Some General Problems Concerning Compilation of Thesauri

Vitězslav Maixner[x]

Введение

Одним из основных теоретических познаний информатики за последние годы является, безусловно, факт, что тезаурусы следует рассматривать как составную часть поисковых языков, причем нельзя упускать из виду значение /роль/ соответствующего грамматического описания. Поэтому было бы неправильным заниматься проблематикой построения тезаурусов, не выяснив некоторых элементарных понятий из типологии поисковых языков.

Однако типология поисковых языков не является главной темой доклада. Поэтому мы ограничимся тем, что приведем несколько рабочих определений из этой области, не претендуя на точность, и затем перейдем к основной проблематике.

Синтаксическо-семантические отношения

Под синтаксисом поиского языка мы здесь подразумеваем любую систему формальных и лексически-формальных средств, четко установленную и общепринятую в описании данного поискового языка /ПЯ/, поскольку эта система дает возможность обозначать семантические /смысловые/ отношения между дескрипторами, составляющими комплексное тематическое описание документа. Для данного ПЯ с синтаксисом существует именно одна такая синтаксичес-

[x] Central Office of Scientific, Technical and Economic Information, Prague.

кая система, помогающая определить множество правильных комплексных тематических описаний на данном языке.

Формальными синтаксически-семантическими средствами ПЯ являются, например, скобки, соединительные знаки, различные знаки препинания, семантически релевантный порядок дескрипторов и т.п.

Лексически-формальными синтаксически-семантическими средствами ПЯ являются, в частности, специальные дескрипторы, соответственно обозначенные в тезаурусе, лексическое определение которых дополнено четкими правилами их синтаксической функции.

Лексические синтаксически-семантические средства неформального типа, будь то целесообразно выбранные или случайные /комбинация дескрипторов для данного случая/ могут изредка полностью или почти однозначно определять отношения между дескрипторами в каком-нибудь конкретном комплексном описании документа. Однако в данном случае не всегда можно говорить о синтаксисе /это не является достаточным условием/.

Использование семантики в ПЯ без синтаксиса, как косвенно следует из предыдущих приблизительных дефиниций и рассуждений, направлено, преимущественно, на системно-лексическую /тезаурус/ и прикладную /индексирование, формулировка поисковых запросов/ области.

При переходе к ПЯ с синтаксисом е с т е с т в е н н а я п р о б л е м а о т н о с и т е л ь н о й о д н о з н а ч н о с т и о п и с а н и я д о к у м е н т а попадает частично в плоскость синтаксическо-семантических отношений /в общих чертах сравнимой с синтаксисом предложения и семантикой предложения естественных языков/.

Проблематика системно-лексического характера /создание тезауруса и его пересмотр/, в основном, присуща поисковым языкам с синтаксисом и без синтаксиса. Однако требования к объему тезауруса /это касается, в первую очередь, количества дескрипторов и определенных лексических сочетаний/ смогут быть, в результате внедрения целесообразного синтаксиса, значительно сокращены.

В некотором смысле можно сказать, что в сравнимых по качеству случаях применения /эффективность поисковой деятельности/ функцию тезауруса перенимает отчасти синтаксис. Синтаксис, лишенный такой компенсационной функции, явился бы неизбежно неэффективным усложнением поискового языка.

Аналогичное взаимнокомпенсирующее отношение можно предположить между функсиями формальных и лексико-формальных средств в синтаксических системах.

В настоящее время существует уже большое количество поисковых языков с синтаксисом, которые отличаются друг от друга хотя бы в формальном отношении. В целях нашего ориентировочного обзора введем сначала следующую общую классификацию, выдвигающую семантические аспекты синтаксиса:

I/ ПЯ выражающие лишь степень взаимосвязи между дескрипторами в комплексном описании документа

2/ ПЯ выражающие лишь функцию /роль/ отдельных дескрипторов в комплексном описании документа

3/ ПЯ выражающие и степень взаимосвязи, и роль дескрипторов

4/ ПЯ выражающие взимосвязь между дескрипторами в комплексном описании документа более явно, чем ПЯ, указаннне в пунктах I-3.

Если мы, наоборот, подчеркнем формальные аспекты синтаксиса, по своему характеру более или менее второстепенные, то можно различить следующие его релевантные знаки с призначными или бесризначными вариантами:

а/ лексичность
б/ позиционная релевантность
в/ формально-символическое обозначение отношений.

Если мы предположим полную сочетаемость призначных и беспризначных вариантов а/, б/ и в/ и учтем комплексным образом семантические и формальные аспекты, то мы легко придем к заключению, что для вышеуказанной ориентировочной классификации можно теоретически образовать 32 различные типовые комбинации ПЯ с синтаксисом. Комбинация исключительно беспризначных вариантов а/, б/ и в/ вгиду данного определения ПЯ с синтаксисом бессмысленна, и поэтому максимельное число типовых комбинаций ПЯ следует сократить до 28. Но количество осуществленных типовых комбинаций, по всей вероят ости, находится под этой теоретической границей.

Семантические проблемы
составления тезаурусов

Литература, занимающаяся методикой составления тезаурусов, сравнительно доступна. Специфичность понятия "тезаурус" различными авторами толкуется по-разному. В нашем докладе мы имеем в виду лишь те тезаурусы, дескрипторы которых находятся в прямой формальной связи с лексическими единицами какого-либо естественного языка.

Общеизвестно, что на практике составление тезаурусов является очень трудоемкой работой, которую можно схематически разделить на три этапа:

1/ подбор источников
2/ составление тезауруса
3/ испытание и пересмотр.

В следующих замечаниях мы попытаемся изложить в общих,теоретических чертах эту методику, с точки зрения семантики ПЯ:

К п. 1/ Собрание терминологических материалов, на котором в конкретных случаях базируются, является, принимая во внимание мировую библиографию терминологических публикаций по данной тематической области, всегда неизбежно случайным и неполным. Влияние этого исходного материала, естественно, очень значительно, хотя и желательно, чтобы оно было подчинено единой концепции построения тезауруса. Наиболее благоприятен был бы, по-видимому, такой ход работ, при котором бы отбор материалов проводился в соответствии с заранее разработанной четкой концепцией.

К п. 2/ На этом этапе отсутствие четкой концепции следует рассматривать как серьезный недостаток.

Разработка такой концепции предусматривает наличие:

1/ грамматического описания ПЯ, по которому можно определить приблизительную типовую характеристику ПЯ и основные правила индексирования и формулировки поисковых запросов;

2/ основных сведений о справочно-информационном фонде - об его объеме, среднем годовом приросте, тематической структуре, о преобладающем методе обработки документов и поисковых запросов

а/ в среднем за несколько последних лет

б/ в перспективном развитии на ближайшие годы.

Концепция построения тезауруса должна содержать:

А/ Определение главных семантических областей тезауруса, иногда также подбор главных лексических дескрипторов /как правило, лишь один дескриптор для определенной области/;

Б/ Диапазоны количества дескрипторов для всего тезауруса и для отдельных семантических областей;

В/ Возможное определение, в соответствии с типом ПЯ, основных формально-лексических дескрипторов;

Г/ Установление всех типов ссылок, которые будут последовательно применяться в тезаурусе;

Д/ Приблизительное определение объема, в каком будут приводиться фразеологические сочетания /соразмерно с количеством дескрипторов/.

Такая концепция, поскольку она приблизительно соответствует предпосылкам I/ и 2/, указанным выше, должна была бы представлять в положительном смысле ограничивающий подбор факторов, облегчающий практическую работу над составлением тезауруса и сигнализирующий возможные ошибки и ошибочные тенденции в ходе этой работы.

Однако в настоящее время об адекватности концепции можно судить, как правило, только "ex post", т.е. после экспериментального испытания тезауруса, что, безусловно, не является идеальным положением. Путь, надежно ведущий от предпосылок к достаточно адекватной концепции, должен был бы быть в максимальной мере подготовлен информационными теоретиками. Но пока нам остается констатировать далеко не радостную действительность, что пока в этом направлении не было в мировом масштабе почти ничего предпринято.

Примечание I. Вышеприведенный методический принцип можно вкратце выразить так: грамматическое описание ПЯ и характеристика СИФ-а приблизительно определяют концепцию тезауруса. В более общей плоскости можно было бы это интерпретировать как некоторое заведомо использование взаимоотношений между грамматическим описанием ПЯ, характеристикой СИФ-а и структурой тезауруса. С формально теоретической стороны, т.е. если не принимать во вни-

мание методические проблемы на практике, можно рассматривать
любой элемент этой троицы в зависимости от остальных двух эле-
ментов, что можно схематически записать так:

    I/ ПН = f/F, T/
    2/ F = g/ПЯ,T/
    3/ T = h/ПЯ,F/.

Мы пока явно рассматривали только отношение 3/, но мы не ут-
верждаем, что отношения I/ и 2/ всегда теряют на практике значе-
ние.

Из отдельных пунктов вышеприведенной схемы концепции состав-
ления тезауруса вместе с предпосылками о ПЯ и СИФ-е вытекает:

из А основная ориентация семантической структуры тезауруса

из Б степень сокращения синонимических кругов исходного естест-
венного языка

из В редукционная обратная связь со всеми главными семантически-
ми областями тезауруса /потенционально обуславливающая бо-
лее высокую среднюю способность комбинаций дескрипторов/

из Г редукционная обратная связь со всеми главными семантичес-
кими областями тезауруса /потенционально обуславливающая у-
точнение местной семантической структуры тезауруса/

из Д обратная связь со способностью комбинаций дескрипторов и
уточнение местной семантической структуры.

    П р и м е ч а н и е   2. В случаях В, Г и Д редукционная
обратная связь может проявляться, в частности, в тенденции к
уменьшению количества дескрипторов /а/или в тенденции к упро-
щению определений дескрипторов.

    П р и м е ч а н и е   3. Уточнение местной семантической
структуры в случаях Г и Д может проявляться, в частности, в до-
полнении определений или небольших семантически связанных групп
с помощью отношений синонимичности, иерархичности или контекста.

Поправки, проводимые в тезаурусе, могут находиться в рам-
ках заранее разработанной концепции или, в крайнем случае,пред-
ставлять существенное изменение первоначальной концепции /пос-
кольку она вообще была ясно сформулирована/.

Не претендуя на полноту, систематичность и единообразие кри-
териев можно различать следующие типы изменений:

    I/ объединение, разделение, исключение или дополнение в си-
стеме основных семантических областей

2/ изменения в иерархической структуре внутри какой-либо основной семантической области /без изменений в соответствующем неупорядоченном множестве форм дескрипторов/

3/ введение или исключение дескрипторов с последующим изменением иерархической структуры.

Экспериментальное испытание функции тезауруса, спорадические исправления, дополнения и запланированные систематические пересмотры всего тезауруса или семантических областей служат одной цели: чтобы семантическая структура лучше соответствовала характеристике фонда в его развитии и данному грамматическому описанию ПЯ. Однако роль тезауруса не должна быть в этом смысле всегда только лишь пассивно приспособляющейся. Стабилизированная или относительно устойчивая структура тезауруса оказывает существенное влияние на некоторые параметры СИФ-а, в частности, на его разделение по различным классам частотности. Можно даже предвидеть такой случай, когда оптимализированная семантическая структура тезауруса вместе с характеристикой СИФ-а будут подсказывать необходимость изменения грамматического описания ПЯ или выбора другого ПЯ /существенное изменение может быть равносильно выбору отличного по типу ПЯ/.

Заключение

Настоящий доклад ставил своей целью вкратце обрисовать некоторые основные проблемы построения тезаурусов с общей точки зрения. Ясно, что многие из этих проблем заслуживали бы более детального специального изучения. Пока в нашем распоряжении не много таких работ, и остается выразить надежду, что теоретики по информатике уделят им в ближайшие годы больше внимания.

Литература:

А.И. Черный: Общая методика построения тезаурусов, НТИ № 5, 1968 г.

O.Seohser, J.Mjžišek, M.Königová: Selekční jazyk a jeho popis. Praha 1968.

D.Soergel: Klassifikationssysteme und Thesauri. Frankfurt/M, 1969.

Р е з ю м е

Тезаурусы, находящиеся в прямой формальной связи с лекси-
ческими единицами какого-либо естественного языка, рассматри-
ваются как частьселекционных языков, типология которых кратко
намечена. В общей плоскости учтены отношения между грамматичес-
ким описанием поиского языка, структурой справочно-информацион-
ного фонда и структурой тезауруса. Приведены некоторые основные
элементы концепции для построения тезаурусов и указано влияние
этой концепции на структуру тезауруса.

S.o m e   G e n e r a l   P r o b l e m s   C o n c e r n i n g
C o m p i l a t i o n   o f   T h e s a u r i

S u m m a r y

Thesauri, based on a direct formal correlation with lexical
units of a certain natural language, are conceived as integral
parts of retrieval languages, the typology of which is briefly
presented. Special attention is devoted to some general aspects
of the relations to be found between a retrieval language gram-
mar, file structure and structure of the thesaurus. Some prin-
ciples directed towards a controlled compilation procedure are
proposed.

STATISTICAL ANALYSIS OF DOCUMENTATION FILES - SADF

Josef Mojžíšek[x]

    Statistical analysis of a documentation file is carried out
by two special computer programmes.
    The first programme is used to test the distribution of do-
cuments in the file as well as exploitation of the file in re-
trieval requests, the aim being to modify the classification
system and thus also the file structure in such a way, that it
may become more suitable for manual storage and retrieval.
    The print-outs of the second program /they are 3 in number/
provide much lucid inforu tion on the documentation file or on
its manageable sample, which is used as input data.
    The processing of documents consists in assigning one or
more simple indexing terms /further SIT/ to each document. The
group of SIT's is called retrieval document description /fur-
ther RDD/ - FIG. 1 A.
    The whole RDD punched on a tape serves as input data for
each document entering the analysis. The computing system is
so arranged as to allow simultaneous processing of up to
10,000 documents. - FIG. 1 B.
    As the result of data processing we get 3 output print-
outs and some other information.
    The first print-out is called a "Table of simple indexing
terms" and it is used as an auxiliary table to ascertain an
unknown rank for a known SIT. - FIG. 2.
    The second print-out is called "Survey of SIT functio-
ning". It is divided into blocks, each block being reserved to
one SIT. Columns 1-10 involve only the given SIT and columns
11-14 show the relation of the given SIT to the   so-called

─────────
[x] UVTEI, Prague.

concurrent SIT's, which are in combination with it. - FIGS.3 and 4.

In this second print-out there is the concluding information:
- average number of RDD's
- average number of concurrent SIT's - FIG. 5.

The last - the third print-out - "RDD Frequency list" - characterizes individual RDD's. It consists of pairs of lines:
- in the upper line there is the RDD written in full - here the RDD is composed of four SIT's
- in the lower line there are 12 columns to be distinguished: See FIG. 6.

The last five columns 8-12 contain ranks of SIT's written in the upper line, which participate in the given RDD.

This 3rd print-out gives the following information:
- average number of documents indexed by one RDD
- average percentage of documents indexed by one RDD
- and average increment of entropy - not yet evaluated - FIG. 7.

A special print-out gives the following very important data:
- number of occurrences of all SIT's
- number of all processed documents
- number of SIT types - FIG. 8.

E v e . u a t i o n   o f   O u t p u t   P r i n t - o u t s

The output print-outs offer us rich, well-arranged information on the document file, or more exactly on the processed sample. They present an objective picture of the structure of the file. The data obtained can be used in many ways, such as for the revision and improvement of the indexing system, for the objective ascertainment of principal and peripheral problems of the file, for the compilation of a thesaurus and so on.

The relation between rank and frequency or accumulative frequency is given by the table in FIG. 11, which is in fact an extract from the "RDD Frequency list" - /3rd print-out - FIG. 6/.

Graphical representation of this table is in FIG. 9.

The axis of ranks is expressed on a logarithmic scale /see interval 1-1703/ so that we can study individual sections of the curve.

The real course of the curves can be seen in FIG. 10, where both scales are linear.

When determining the relation of the length of RDD's to their frequency /i.e. to the size of equivalence classes/ we shall work with a new notion of an iso-frequency group.

In order to determine this relation the table in FIG.12 was compiled.

Each iso-frequency group is identified by two characteristics:

- by the RDD frequency of this group - for instance 24 /framed/
- and by the interval of the RDD ranks belonging to this iso-frequency group.

The purpose of the method is the analysis of RDD's by their length, i.e. we try to ascertain how many SIT's compose RDD's belonging to individual iso-frequency groups.

The diagram in FIG. 13 graphically presents the left part of the table in FIG. 12. Each iso-frequency group is divided, in a linear way, into subgroups by different RDD lengths, i.e. by the number of SIT's involved in the RDD.

The diagram indicates how many iso-frequency groups exist, the number of RDD s included in each group, and frequencies of individual RDD's as well as their distribution by length in aach iso-frequency group.

The values in the right-hand part of the table in FIG. 12 are the basis for plotting the diagram in FIG. 15, which is a graphical representation of analysis regarding RDD frequencies. This diagram demonstrates the relative parts of the file included in separate iso-frequency groups as well as the distribution of these groups by the length of RDD's. It can be used as a nomogram. It presents, also, how the file is distributed into RDD's by their length.

The analysis, the result of which indicates the combination power of indexing terms, starts from blocks for individual

SIT's - we process columns 8 to 14 of the second print-out
"Survey of SIT Functioning". - FIG. 3.

Then in the third print-out we gradually find RDD's with
ranks given in front of horizontal lines - examining columns 8
to 12 and marking by circles those SIT's that are participa-
ting in particular RDD's.

As an example the following block of SIT with rank 73 -
FIG. 21 - was proceased.

A graphical display is to be found in diagrams in FIGS.22
to 24.

These investigations will allow us to learn a great deal
about related terms or descriptors /SIT's/.

It is probable that the needs of practice would lead to
the alternation of programme to the end that it might offer
the maximum of useful information in the most convenient form.
The proposed methods of the second programme regard only the
analysis of a document file, although we know that the program-
me could also be used for the evaluation of the set of retrie-
val requests.

NOTE
All Figures see in the Annexes

COMPILATION OF THESAURI FOR USE IN COMPUTER SYSTEMS

Loll N. Rolling[x]

I - Introduction

A thesaurus can be defined as a structured vocabulary for
use in information storage and retrieval systems.
Three parts of this definition need further elaboration:
1. A vocabulary is a collection of terms.
2. The structure of a vocabulary can be described as a set of
   relationships between terms.
3. Utilization of a thesaurus in an information system invol-
   ves a set of rules which take into account the characteris-
   tics of the system.

x

1. There are three types of thesauri according to the type
of terms they consist of.
A few of the earlier thesauri consisted solely of uniterms[1],
i.e. single words. Some of these acquired significance only in
combinations. Even simple concepts had to be represented by a
combination of uniterms.
Many thesauri are of the "uniconcept" type. Uniconcepts can
be either uniterms or polyterms, i.e. single words or combina-
tions representing simple concepts.
Frequently co-occurring uniconcepts and uniterms can be
combined to form pre-coordinated terms of the "subject heading"
type. Most of today's thesauri comprise both uniconcepts and

[x] European Community, Luxembourg

phrases representing a pre-coordination of two or more concepts,
which could be called "polyconcepts".

It is generally accepted that the vocabulary of a thesaurus
should be homogeneous. However, there are a few exceptions. The
thesauri of the USAEC[2/] and MEDLARS[3/],for example, which are
generally of the "subject heading" type, comprise a limited
number of uniterms /called "subheadings"/ to be used only as role
indicators as it were in   combination with the main headings.

                                  x

2. There are three classes of relationships between the terms
of a thesaurus.

The "optional" or "indicative" type is represented by the
crossreference   _A see also B_, which invites the indexer   in
the process of assigning term A to _see_ whether term B is not
also relevant.

Also of the optional type are the RT /related term/, NT
/narrower term/ and BT /broader term/ relators now in use in
most of the English-language thesauri. The NT relator invites
the indexer to check whether he should not be more specific.The
BT relator suggests that the concept to be indexed might have a
wider coverage than the main entry consulted.

The "compulsory" type is represented by the reference _A use
B_. In many thesauri this means that term A must not be used
and term B assigned instead /"exclusive" reference/. Term.B is
then either a preferred synonym or an abbreviation, or a spelled
out version of term A. But it could also mean that term B should
be assigned in addition to term A /"complementary" reference/.
Such "complementary" references are mostly "generic" references
from a specific term to a term of a higher hierarchical level,
and the additional assignment of term B is called _generic pos-
ting_.

In a thesaurus where exclusive and complementary references
co-occur, it is ne:essary to tag them with dif~erent symbols or
relators.

A use B must be distinct from _A use also B_ or _A add B_

A third class of relators is necessary to represent the
"alternative" relationships which exist between the various mea-
nings of homographic terms.

Homographic terms cannot be allowed in a retrieval system
if selection of non-relevant information is to be avoided.

The references from a homographic term to the alternatives
offered must therefore be of the "exclusive" type.

A see B or C leaves the indexer free to ch'ose which
term he will assign to replace term A.

In a number of thesauri the difficulty is overcome by ad-
ding more or less elaborate specifications to homographic terms:
REPRODUCTION /BIOLOGY/ is unmistakably differentiated from RE-
PRODUCTION /COPYING/[4].

One of the disadvantages of this method is that the inver-
ted form of composite words has to be maintained in the alpha-
betic list if the alternative terms are to be found. CONDUCTI-
VITY /ELECTRIC/ and CONDUCTIVITY /THERMAL/ can be found by an
indexer looking for CONDUCTIVITY, but ELECTRIC CONDUCTIVITY and
THERMAL CONDUCTIVITY cannot.

Another solution consists in ignoring all but one of the
meanings of a homographic term: FRONTS /METEOROLOGY/ rules out
all other meanings of the word FRONT and other terms such as
FOREHEAD, FRONTAGE, and BATTLE-LINE must be resorted to to ex-
press these concepts. In some thesauri the indications limi-
ting the use of a number of terms to one of their meanings are
extended in the form of lengthy scope notes.

x

3. Structured vocabularies can be used in

a/ conventional information systems

b/ computer-assisted systems and

c/ fully automatic, interactive systems.

In conventional systems the document file can be arranged
by thesaurus terms; there must be as many document copies in
the file as thesaurus terms assigned to the document.

If the document collection is arranged by author name,for-
mat or in chronological order, a separate index file is prepa-
red which contains, in the order given by the thesaurus /gene-
rally alphabetical/, one abstract or title card for each the-
saurus term assigned to each document.

In a "dual-file" system, the abstract /or title/ cards are
in numerical order /direct file/, while a second file contains

one card for each thesaurus term with an indication of the do-
cument numbers to which the term was assigned /inverted file/.

The most popular dual-file system is the "peek-a-boo card"
file.

Optional and alternative references can be noted on insert
cards; compulsory references can be implemented by introducing
additional file cards or punching additional peek-holes cor-
responding to the generic terms.

In computer-assisted and fully automatic systems, the the-
saurus terms and structures as well as the files are stored on
magnetic /or, exceptionally, photographic/ media.

The use of the computer's sort and print facilities for
storing, sorting, updating, and printing successive versions
of a thesaurus in various presentations is very popular. Compu-
ter-printed alphabetical lists, permuted term lists, inverted
term lists, subject category lists, etc., can be found in many
English-language thesauri.

The storage of the generic and semantic relationships be-
tween terms with a view to their application to stored indexes
is a relatively new technique. This will be discussed in Chap-
ter II.

The computer storage, processing and retrieval of direct
and inverted files, again, is practised in all non-conventio-
nal systems, with widely varying results. These results, as
shown in Chapter III, are largely conditioned by thesaurus
structure and control.

II - Computer   manipulation
of   term   relationships

Preparing a retrieval file amounts to compiling the maximum
number of relevant thesaurus terms as access points for retrie-
val by term coordination. The problem is one of identifying
those thesaurus terms which are explicitly or implicitly con-
tained in the tape-stored starting material. The starting mate-
rial can be a corpus of titles, abstracts or full document
text, or alternatively a set of manually assigned index terms.

In the first case, the computer is fed strings of non-standard, but generally correctly spelled words, and the "automatic indexing" routine will result in the selection of formally correct terms, the relevance of which cannot be guaranteed. The routine will involve splitting the text into single words for the identification of thesaurus uniterms, and into groups of two or more words for the identification of composite thesaurus terms. The identification of terms is complicated by the presence of non-standard word forms involving prefixes and suffixes, and by the occasional insertion of non-significant words between term components.

In the second case, the computer has to deal with generally standard, but often incorrectly spelled individual index terms, of guaranteed relevance, which can be directly matched with the thesaurus.

In both cases the computer can then be used to perform "generic posting", i.e. to carry out the information transfer suggested by compulsory references of the _A add B_ type.

Note that B does not have to be an indexing term. In the Euratom system, for example, there are several hundred _A add B_ references for the posting on to terms of a higher generic level, such as names of categories and disciplines, and cumulative designations which it would be nonsensical to use for the indexing of documents, but which are quite selective in the retrieval process.

The posting on to generic and cumulative terms does away with the need to include long arrays of alternative terms in query formulations.

Contrary to what might be expected, the expense occasioned by expanding the retrieval file through intensive generic posting is lower than the additional cost of manipulating strings of alternatives in the retrieval process.

Generic posting can be combined with a correction routine based on the _A use B_ references. The correction routine amounts to replacing the incorrectly assigned term A by its preferred synonym B.

In the Euratom system this procedure was extended to encompass the correction of spelling errors introduced by indexers and keypunch operators. The errors detected by thesaurus

matching and corrected manually can be fed into an "error dic-
tionary" the format of which resembles that of the thesaurus.
E use A means that every time the thesaurus term A is again
misspelled E, it will be automatically corrected to A.

Among the errors occurring for the first time, those pro-
duced by addition or omission of one character, substitution
of one character for another, or inversion of two adjacent cha-
racters, can be corrected by a separate computer programme,
which changes, for example, ALCOOIS, ALCOHOLES, ANCOHOIS and
ALOCHOIS into ALCOHOIS.

Every error corrected by this programme can in turn be fed
into the "error dictionary" as an additional E us3 A referen-
ce.

The correction routine using term and error dictionaries
can be developed, by inclusion of all possible word forms and
spellings and a list of non-significant words, into a programme
for automatic indexing. Experience has shown this approach to
yield better performances than the method using word stems and
truncated terms[6].

However, the quality obtainable by automatic indexing is
still largely inferior to that of indexing by specialists, and
it can only become attractive if this shortcoming is balanced
by lower cost. The major cost component being that for input
rather than computer processing, automatic indexing will spread
as the availability on tape of the starting material, i.e. the
texts of documents and abstracts, comes into general use as a
byproduct of publication. Optical character - recognition devices
are not expected to become economic to the point where automa-
tic indexing would be competitive.

The computer is also used to check the internal consisten-
cy of the thesaurus structure. In particular, every time a new
term or a new reference is introduced, a complex automatic rou-
tine checks that the introduction does not lead to duplica-
tions, contradictions or continuous loops within the thesaurus
or between the thesaurus and the error dictionary.

III - Effect of generic posting
on thesaurus building

The number of A use B references in a thesaurus is ge-
nerally kept as low as possible: It seems senseless to over-
burden a term list with terms that the indexer is not permit-
ted to use. The majority of the terms are thus .authorized
for indexing and retrieval. This makes it easy for the indexer
to find at least one term corresponding to tne concept he has
in mind, but the retriever has to browse through a great many
RT, NT and BT references in order to assemble all the terms
relevant to the query, i.e. all the terms that the indexer
could have assigned to relevant items. The use of automatic
generic posting radically alters the picture. First,the inde-
xer cau now limit himself to assigning the most specific terms
corresponding to the concept to be indexed, since every rele-
vant generic term will be automatically added. Indexing depth
will decrease, as also the indexer's workload per item.On the
other hand, the retriever will not have to bother about inclu-
sion of alternate terms, since every term he may select for
query formulation will bear the postings of all hierarchically
subordinate terms.

In the Euratom System, where generic posting is being app-
lied to a very large extent, the influence on retrieval per-
formance has been the following:

Relevance ratio has significantly increased, since the
indexers are free to follow their natural inclination, which
is to put syatematically more emphasis on specific indexing.

Recall ratio also increased, since it is no longer possible
for the retriever to overlook relevant specific terms in the
query formulation.

As both indexing and query formulation have become elemen-
tary operations, they take less time and have become less
costly.

The Euratom Thesaurus was c mpiled as early as 1962 as a
set of equivalent descriptors, ollowing the example of
ASTIA[8].

A great many A use B references were then developed as
result of the indexing of more than 100,000 abstracts, and

the indexers were systematically invited "to use term B in
preference to term A whenever pertinent".

When, in 1964, Euratom's first computer was used to verify
consistency of application of this rule, it became evident that
the same inexpensive operation could be used for correcting
errors /B use A/ and for generic posting /A add B/.Throughout
the following years, _A add B_ references from new specific
terms /A/ to already existing descriptors /B/ were added to the
thesaurus every time it was felt desirable. The distribution
of the new references over the subject field was therefore not
uniform. And the descriptor part of the thesaurus remained vir-
tually unchanged, since many of the modifications which might
have been considered desirable would have necessitated exten-
sive re-indexing.

The situation is quite different in the field of metallur-
gy.

The European Community is now considering extending its
information dissemination activities into the fields of metal-
lurgy and, eventually, agriculture.

Preparatory work is being done including an inventory of
the metallurgical literature and compilation of a metallurgi-
cal thesaurus.

This seems an excellent opportunity to apply the new meth-
odology of thesaurus building.

A number of basic rules and a plan of work were agreed
upon, and work on the thesaurus started in October 1968.

The following is an outline of our plan of work and the
way we went about it.

S t e p   1: Define the subject field to be covered and
make an inventory of existing thesauri and previous terminolo-
gy studies.

R e s u l t: Three English-language thesauri compiled by
ASM[9], CSM[10], and ISCOR[11]; a French thesaurus developed
by CNRS[12]; a number of excellent encyclopedias; the Metals
Handbook; and the subject indexes of the major abstract jour-
nals.

In view of the number and quality of these documents it
was considered superfluous to compile a representative set
of terms by statistical analysis of a text corpus.

S t e p   2: Eliminate duplicates from the term collection.

R e s u l t: From a total of 10,000 terms originating from
three thesauri, 1,100 duplicates and 900 triplicates were dele-
ted by alphabetical merging of the term lists, leaving a collec-
tion of 8,000 terms.

This step could have been combined with step 6; taken sepa-
rately, it made steps 3 and 5 easier.

S t e p   3: Divide the subject field into coherent units
containing 100 to 300 terms.

R e s u l t: Creation of 37 term subsets covering various
aspects of metallurgy, including materials, properties and pro-
cesses.

S t e p   4: In heavily loaded subsets, take special measu-
res to standardize term selection and format.

R e s u l t: Creation of a rule for the generation of de-
signations of inorganic compounds, ores, alloys and isotopes,
by combination of element names with the words COMPOUNDS, ORES,
ALLOYS, ISOTOPES, etc.

This rule may well burden the printed thesaurus, but not
the user's memory.

S t e p   5: Display the terms corresponding to each unit
in semantic charts, grouping conceptually related terms around
their preferred synonyms.

R e s u l t: 900 preferred terms, defined by their semantic
context, earmarked for descriptor status, with an average of 8
terms clustered around.

S t e p   6: Define the reference type for each term and
record the references for computer storage.

R e s u l t: 3,800 descriptors, 1,000 A use B references,
50 A see B1 or B2 references, and several thousands of A add B
references, on 80-column thesaurus form sheets.

Almost every term had to be verified in one or more hand-
books and encyclopedias, decisions about homographic terms being
the most time-consuming.

A number of terms were considered valueless in a thesaurus
for use in a system based on term coordination, but they were
retained, generally as A use B references, in order to achieve
complete convertibility between the basic term lists/9/ /10/11/
and the new thesaurus.

S t e p   7: Computer storage, consistency check and print-
out.

R e s u l t: All the terms keypunched on 80-column cards,
stored on magnetic tape, sorted alphabetically by first and by
second terms, resulting in a printed thesaurus and an "inverted
dictionary" with related terms grouped in synonym clusters.

Duplicate entries, contradictory entries and references
loops are eliminated in the process.

S t e p   8: General revision and verification.

R e s u l t: The display charts were checked against the
"inverted dictionary" and given their final presentation.

Term lists and display charts were critically examined by
several experts; as a result of their comments a number of mo-
difications were made.

The total effort involved in the compilation of the metal-
lurgical thesaurus by two members of the staff and a small
number of experts participating in the project was calculated
at approximately 700 working hours, or four man/months.

Production of the two listings required only a few minutes
of computer time using the IBM-360/40.

The above figures seem quite incredible when compared,for
example, with the tremendous effort that went into the EJC[5]
or DoD[4] Thesauri. Wall's figures[7],[13] represent a total
of 70 man/months.

It is not surprising that, in the light of such an enorm-
ous expenditure of expert time /and money/ many studies. were
started on the possibilities of automatic generation of the-
sauri[14], which by the way, in turn, are costing a lot of
computer time /and money/.

Now that the efficiency of automatic posting has been con-
clusively demonstrated, and with graphic display of semantic
and hierarchic relationships making the compilation of generic
references such an elementary operation, there is no doubt that
more and more thesauri will be built, economically and effi-
ciently, u   .g these methods.

IV - S i m i l a r i t y   f a c t o r s

Generic posting takes care of the BT references in indexing
and their NT counterparts in retrieval. The RT relators,however,

representing a variety of ill-defined relationships, have esca-
ped all attempts at computerizing so far. The documentalist in
charge of indexing or query formulation will assess, from the
display of semantic relationships in the terminology charts,
which is the term most likely to represent a valuable alterna-
tive to a given concept because of its similarity or closeness to
it. If this "semantic similarity" between two concepts or terms
could be given a numerical value, the computer could be taught
to store and use semantic similarity factors for a number of
/or all/ term couples in a thesaurus.

The expression A/50/B might be used to mean that B is
semantically similar to A to the extent of 50 percent, and a do-
cument tagged with term B might be expected to be half as rele-
vant as a document tagged with term A.

Various methods have been thought up, aiming at making the
computer determine statistically the value of the similarity
factors on the basis of an indexed data base[15]. So far these
experiments have not been conclusive.

We therefore consider making use of numbers assigned on
the basis of expert judgment in the planned experiments invol-
ving similarity factors.

The basic experiment, of which a number of variations are
under consideration, can be described as follows:

A question expressed by the customer as a combination of
three aspects: A   B   C   would normally be modified by the
information officer to include a number of alternative terms
/A2, B2, B3/ which he judges to be relevant in this particular
case. The query thus becomes

$$/A + A2/ \quad /B + B2 + B3/ \quad C$$

If, however, the computer knows the similarity factors of
all term couples involving A, B, and C, it can be told to se-
lect all terms having a similarity above a given cut-off value,
say, 50 percent, and apply these as coefficients to the terms
in the Boolean formula:

$$/A + 0,8 \ A2/ \quad /B + 0,9 \ B2 + 0,6 \ B3/ \quad C$$

A document tagged /A, B, C/ would be retrieved with a re-
levance probability /the product of the similarity factors in-

74

volved/ of 100%, e document tagged /A2, B, C/ with a probability of 80%,
a document indexed /A2, B2, C/ with a probability of 72%, and a document
indexed /A, B3, C/ with a probability of 60%.

The computer, using the similarity factors, thus not only retrieves
all the documents found by the human querist, but produces a ranking of
the selected references according to their relevance probability. This
relieves the system operator /or the final user/ of part of his burden and
reduces the total cost of information retrieval.

Similarity factors are expected to play a rôle in the man-machine
interactive information systems of the future.

References

1/ M.TAUBE et al.: Studies in Coordinate Indexing, 5 vol.,Documentation
   Inc., 1953-59.
2/ USAEC: Subject Headings used in the catalogs of the United States
   Atomic Energy Commission, /TID-5001/, 8th revised edition, Jan. 1969.
3/ National Library of Medicine: Medical Subject Headings. Jan. 1965.
4/ U.S. Department of Defense: Thesaurus of Engineering and Scientific
   Terms /AD 672000/ 1967.
5/ Engineers Joint Council: Thesaurus of Engineering Terms. May 1964.
6/ E.M. KEEN: Thesaurus, Phrase and Hierarchy Dictionaries. Part VII of
   Cornell University. Report ISR-13: Information Storage and Retrieval.
   Jan. 1968.
7/ U.S. Department of Defense: The Making of an Interdisciplinary The-
   saurus. /Chapter 1 of ref. 4/.
8/ Armed Services Technical Information Agency: Thesaurus of ASTIA
   Descriptors. 1st edition - May 1960.
9/ American Society for Matals: ASM Theseurus of Metallurgical Terms.
   Dec. 1966.
10/ Centro Sperimentale Metallurgico: Thesaurus per la Siderurgia.Edizione
    anglo-italiana. May 1966.
11/ South African Iron end Steel Corporation: ISCOR Thesaurus, 2nd edition,
    Jan. 1968.
12/ Centre National de la Recherche Scientifique: Thesaurus de Chimie
    Appliquée - Métellurgie. 1968.
13/ E. WALL: Vocabulary Building and Control Techniques. American Documen-
    tation. Vol. 20, pp. 161-64. April 1969.
14/ C.D. BATTY: The Automatic Generation of Index Languages. Journal of
    Documentation. Vol. 25, n° 2, pp. 142-51. June 1969.
15/ G. LUSTIG: Automatic indexing as a complement to the semiautomatic
    documentation system. Part 2 of report EUR-4256   3.

ANSWERS TO THE QUESTIONNAIRE ON THESAURUS PROBLEMS

Grete A.Schanche[x]

1a A thesaurus always comprises an alphabetically arranged,
    and may also contain a systematically arranged display of
    evaluated subject-oriented terms for indexing and retrie-
    val purposes.

1b To call a given construction a "thesaurus" the following
    semantic and syntactic principles have to be fulfilled:
    a/ The thesaurus has to be structured as to synonyms and
    quasi-synonyms, for instance by means of USE-references,
    SEE ALSO-references and USED FOR-references.
    b/ The thesaurus must contain definitions of ambiguous
       terms, for instance by means of scope notes added to
       the term.
    c/ The thesaurus must be accompanied by a set of rules
       giving instructions for:
       use of the various grammatical forms of the words,like
       the use of nouns, adjectives, adjectives together with
       nouns, use of singular-plural form and so on,

       use of accepted abbreviations /like "VTOL aircraft"
       for "vertical take off and landing aircraft"/,

       use of symbols like - , //, greek letters, "per cent",
       "degrees centigrade" etc.,

       use of trade names,

       use of geographic names,

       use of technical slang,

---

[x] Studieselskapet for Norsk Industri, Oslo.

combination of single terms in the thesaurus to complex
terms by indexing and retrieval,

direct or inverted entry of complex concepte consisting
of more than one word /like "fluidized bed" and "bed
fluidized"/,

the candidate terms must be evaluated by e subject epe-
cielist ae well as by e linguist.

The abovementioned are the basic principles of thesaurus
construction. In addition to these there are other prin-
ciples - the fulfillment of which give rewarding advan-
tages to the thesaurus ae a tool for indexing and   re-
trieval:

A structurization of the thesaurus by means of a hierar-
chical listing of the terms or by a listing of broader
term /BT/ - narrower term /NT/ reletionships between the
terms.

Such a structurization will cause secondary practical
effecte on the thesaurus by giving possibilities   for
preparing alphabetical lists for groupe and subgroupe
within special fields. These may, besides serving their
main purpose /indexing and retrieval tool/, be of value
for studying the depth of indexing by means of word fre-
quency studies. Such studies will in turn contribute to
the development of the thesaurus towards an optimaliza-
tion of usefulness.This means that the thesaurus   con-
tains a minimum number of rarely-used or too-often-used
terms and a maximum number of relevant terms. It means
further that the thesaurus should be supplied conti-
nuously with new evaluated terms and should be freed
from obsolete terms by continuous   deletion. This last
point is in accordance with rapid technical development
within the various special fields.

1c The elements or factors which influence the construction
of a thesaurus are mentioned in 1b. Besides these factors,
the aim of the thesaurus ought to be considered from case
to case; whether it is going to serve a "polytechnical"
/several branches/ or a more specialized field, whether it

is going to serve one branch /company/ or one subject field
and whether it is going to serve as a tool for both index-
ing and retrieval purposes. It is also important to evalua-
te the desired depth of indexing from case to case. This
in turn will influence the thesaurus construction.

1d The degree of complexity and the number of information
    items /terms and various presentation forms of these/ have
    to be evaluated in connection with the aim of the thesaurus
    /answered in 1c/, in terms of what purposes are to be ful-
    filled. The economy of the thesaurus project in question is
    of course also an important factor, which will determine
    the degree of complexity.

2. When applied to the use of the thesaurus for coordinate in-
    dexing and retrieval, the existing definitions found in
    recent literature and used among documentalists /something
    like the definition given above under 1a/ should be suffi-
    cient. If the thesaurus concept is to cover broader uses,
    its definition has to submit to further analysis.

3. The role of the thesaurus is, according to the answer under
    2, primarily for direct use as a most necessary tool in
    systems for indexing and retrieval. Secondarily it can be
    used as a vocabulary and dictionary aid for other purposes,
    for instance within scientific information as a whole,such
    as for authors editing reports, journal articles, for lib-
    rarians in their choice of keywords for all kinds of scien-
    tific library material, etc.

4. The most rewarding way of collecting terms for a thesaurus
    is no doubt the compilation of evaluated terms from current
    indexing of new literature in the field in question, be-
    cause this gives least "noise" with regard to artificially
    and seldom-used terms. It gives the current up-to-date
    language. In this work there are obvious advantages in using
    a computer when it comes to alphabetization of the word
    collection, for preparing alphabetized subject field lists,
    for construction of hierarhical connections,for production
    of inverted lists, permutaved   lists, for word frequency
    studies etc.

5. Of the topics mentioned in the questionnaire the first is
   in our opinion the most important, i.e. the classifica-
   tion of thesauri according to branches/subject. In fact
   both of these classifications can be of value.
   A company thesaurus usually ought to be a "branch" the-
   saurus, while a thesaurus for a scientific institution
   ought to be more subject-oriented.

6. A descriptor is a preferred keyword /term/ which in a
   semantic hierarchy has a place which may comprise a larger
   concept field than the one which is covered by one single
   keyword /term/.

7. The answer to this question is given under 1b.

8. The answer to this question is given under the last part
   of 1b.

RECOMMENDATIONS FOR THE BUILDING OF THESAURI
IN SCANDINAVIAN LANGUAGES

Regler for bygging av thesauri pa nordiske sprak[x]

E x t r a c t   i n   t r a n s l a t i o n
Henning Spang-Hanssen[xx]

As to general points of view and as to order of presenta-
tion, the recommendations are in accordance with the Thesaurus,
Rules and Conventions found as Appendix 1 to the Engineers
Joint Council's Thesaurus of Engineering and Scientific Terms
/New York 1967/.

D e f i n i t i o n s:
By a _document_ is meant the smallest bibliographic unit rel-
evant to a system of registration and retrieval.
By a _keyword_ is meant any word or composition of words ap-
propriate in retrieval for the characterization of the content
of a document.
By a _descriptor_ is meant /1/ a preferred keyword that /2/
within a conceptual /semantic/ hierarchy takes up a position
that /3/ may cover a conceptual field larger than the field
covered by an individual keyword.

P u r p o s e   o f   a   t h e s a u r u s:
A thesaurus serves the purpose of indicating the _transfer_

80

- 80 -

of keywords /found in documents and in retrieval queries/ into
descriptors accepted by the information store of a system of
registration and retrieval.

   Contents   of   a   thesaurus:
   A thesaurus obligatorily comprises an alphabetized list.
This list should include all keywords and all descriptors.
   In addition, a thesaurus may comprise
1/ a permuted index /allowing to enter from inversions of nor-
   mal word order, e.g. from absorption acoustic/;
2/ a subject category index, in which the descriptors are ar-
   ranged according to categories /facets/;
3/ a hierarchical index in which descriptors having BT-or NT-
   references /cf. C-1/ are arranged hierarchically below the
   descriptor having the broadest meaning /covering the bro-
   adest conceptual field/;
4/ a list of notations /e.g. numeric codes/ for descriptors
   and for corresponding keywords.

   Selection   of   descriptors: T-1
   Descriptors are selected in accordance with their use in
documents for indexing. In addition to all conceptually impor-
tant words, forming the core of a thesaurus and being closely
affiliated to professional terminology, descriptors may include
names of persons, institutions, projects, locations, etc., and
furthermore numbers, numerical intervals, and other kinds of
identificatory symbols. Even bibliographical indications may
act as descriptors.
   The usefulness of a descriptor depends on
/a/ whether it is likely to occur in retrieval queries as an
   element conveying information;
/b/ whether it can be given a meaning distinguishable from the
   meanings of other descriptors;
                              Recomm. ... Scandinavian lang. -2
/c/ whether it can be given a definable meaning of general un-
   derstandability.
Not useful as descriptors are words
/a/ never occurring in the relevant documents;
/b/ occurring in any document;

81

/c/ never occurring in retrieval queries;
/d/ having a vague and undefinable meaning;
/e/ having a meaning very close to that of some already-accep-
    ted descriptor.

    For descriptors the S i n g u l a r   o f   n o u n s   is
used.[x]   T-3
    Instead of distinguishing in the way of English building
and buildings, indicators of content should be used when neces-
sary, e.g. building /process/ versus building /object/.

    I n d i c a t o r s   o f   c o n t e n t: T-5
    In cases of doubt as to the meaning of a word to be accep-
ted as a descriptor, the meaning should be pointed out by ad-
ding an indicator of content /a scope note/. The indicator is
given in brackets, but forms part of the descriptor. The fol-
lowing cases deserve mention:
/1/ Cases of homonymy, e.g. English /mercury /metal/ and Mercury
    /planet/.
/2/ Cases where a noun is commonly used both of a process and
    of a product or a property /cf. T-3 above/.
/3/ A composition of words may have different meanings according
    to varying semantic relations between the components, and an
    indicator of content may thus be desirable, e.g.   English
    water cooling /cooling with water/ versus water cooling
    /cooling of water/.
/4/ Trade marks and possibly other names may need an indication
    of the kind of article in question and/or of the fact that
    the name in question is a registered trade mark.

    C o m p l e x   d e s c r i p t o r s   /p r e - c o o r d -
i n a t e d   d e s c r i p t o r s/:   T-11
    Usually descriptors should be chosen in accordance with the
terminology found in the relevant literature regardless of the
number of words used in order to express the concept/the meaning/

---

[x] The rules recommended here for Scandinavian languages thus
    deviate from the American rules.

in question. However, many potential descriptors seem to express concepts that are combinations of two or more other /potential/ descriptors. In such cases a decision must be made as to the inclusion of the specific, complex descriptor or the treatment as a combination of descriptors.

Specific, complex descriptors often facilitate the retrieval of specific information, but increasing the number of descriptors generally increases costs of indexing. The use of individual /not complex/ descriptors by indexing, and combining them during retrieval, permits a smaller thesaurus and of a more consistent terminology.

                    Recomm. .... Scandinavian languages -3

/a/ A specific, complex descriptor should be set up, if relevant and more general descriptors are not found in the thesaurus in question. In order to be relevant a combination of more general descriptors must comprise at least one descriptor that is a member of the same hierarchical class as the specific concept.

/b/ A specific, complex descriptor should be set up, if the specific concept is itself frequently met with,or if one /or both/ of the more general descriptors is very frequently met with.

In cases of doubt it is advisable to set up the specific, complex descriptor, at least temporarily, because in a working system it is easier to split up a complex descriptor than to combine existing descriptors into a new, complex descriptor.

Cross   references: C-1

The connections between the descriptors and keywords within a thesaurus are indicated by 5 reference symbols having fixed meanings. In thesauri in the English language the most widespread symbols are /abbreviations of/ English words

   USE   UF /used for/   BT /broader term/ NT /narrower term/
   RT /related term/.

In a thesaurus in a Scandinavian language it is not recommended to introduce abbreviations specific to this language, because confusion may arise. It is advisable /1/ either to accept the widespread English abbreviations, regarded as pure symbols, or /2/ to standardize - if possible - new, figurative symbols, e.g.

- 83 -

	>	<	▲	▼	//
for	USE	UF	BT	NT	RT

where A and V are to be read as letter substitutes for arrows
pointing respectively upwards and downwards.

/More detailed definitions and instructions with examp-
les for the 5 types of reference are given in the original as
sections C-2 to C-9. These, and rules for alphabetizing /sec-
tion A-1/ are omitted in this extract; from the numbering it
will be seen that even sections T-2, T-4, T-6 to T-10 have
been omitted. This translated extract tends to cover points
of particular interest as to principles of thesaurus construc-
tion/.

PROBLEMS OF THESAURI

Jiří Toman[x]

The questionnaire does not contain questions which most
probably could incite the participants to discussion - namely
questions about the problems and the disadvantages of these-
uri. As the host has invited us to add freely new points to
the questionnaire, this paper tries to concentrate especially
on both these aspects. In order to keep to the order of items
as shown in the questionnaire, I shall deal with the indivi-
dual points in the order of the questionnaire items.

ad 1. - 2.
One of the most important activities of librarians and
information specialists is to index or classify the informa-
tion materials, i.e. to express the topic of documents briefly
by means of terms /called subject headings, uniterms, des-
criptors, key-words etc./. The terms representing the topics
of documents can be arranged either alphabetically- according
to the external form of the term - or systematically - accor-
ding to their internal content. There is no third way of ar-
ranging them. These two ways of arranging the terms corres-
pond to the two main trends of ordering systems, namely -
alphabetic indexing systems /like subject headings, uniterms,
thesauri of descriptors, key-words/ and systematic classifi-
cation systems /both traditional and faceted/.
In English there unfortunately does not exist a broad
term for the two narrow terms "indexing" and "classification".
In Csech we use the broad term of "ordering systems" /simi-
larly the Germans use the expression "Ordnungssysteme"/.

[x] Czechoslovak Academy of Sciences, Prague.

In this paper the words "ordering systems" will be used to
express both alphabetical and classification systems.

The thesaurus is an alphabetical system. It is necessary
to stress this, because in some cases this expression is used
nowadays to describe any special ordering system /as distinct
from a universal classification system/, even if it is syste-
matic. Unclear terminology would exclude discussion; therefo-
re by "thesaurus" we should always understand an alphabetical
system.

ad 3. Role

In conventional information systems it is immediately
evident whether the catalogue is alphabetical or systematic,
because the records in the files are arranged in one or the
other way. In a mechanised system we cannot judge the charac-
ter of the ordering system according to the arrangement of
records. Most frequently their arrangement on magnetic tape
is chronological.

Indexing and classification
fulfil an important role -
namely to arrange the
records in storage

Every classification system uses systematic tables /syste-
matical display of terms/ and an alphabetical index /alphabe-
tical survey of terms/. Which of the alphabetical ordering sy-
stems, whether subject headings, uniterms or thesaurus, uses
both of these aids? I somehow doubt that graphic maps of terms
in a thesaurus correspond to the tables of a classification
system. The lack of a systematic survey is a great disadvan-
tage of many thesauri.

ad 4. Construction

Referring to my papers presented at the Elsinore confe-
rence in 1964 and at the Second Anglo-Czechoslovak Conferen-
ce of Information Specialists in London in 1967 I shall repeat
briefly the main ideas of the or ering principles and the
theoretical considerations connected with it:

Every ordering system is governed by a combination of different ordering principles, whether alphabetical or systematic.

Nearly every one of these principles is in contrast with another principle:

P r i n c i p l e	C o n t r a s t i n g p r i n c i p l e
alphabetical arrangement of terms	systematical arrangement of terms
uncontrolled dictionary	controlled dictionary
no hierarchy	hierarchy
open-ended system	closed system
precoordination	postcoordination
alphabetical display of terms	systematical display of terms /tables/

In addition to this there are two principles which have no antitheses - the principle of categories and the principle of expressing terms by notation. Both these principles are characteristic of the classification systems.

The chief idea expressed in the paper presented in London was, that there was a general tendency of these principles to merge and that the best solution seemed to be achieved by the synthesis of opposing principles:
Uniterms have no hierarchy, while UDC uses a strong hierarchy. Neither an excess nor a lack of hierarchy are good. A weak hierarchy therefore seems to be the best solution.

An uncontrolled vocabulary has great disadvantages, but similarly a controlled vocabulary - as we all feel - very often has the disadvantage of a Procrustean bed. In my opinion there will be a general tendency to add uncontrolled words to the records. I call them "explicative words" and they have no selective power /no search can be made on their basis/. We have introduced these words in our INDORES mechanized information system, and they serve us well. They are very important in the computerized systems where bibliography has replaced abstracts and wher. the strict use of a controlled vocabulary endangers the understanding of the content of a document.

Precoordination or postcoordination - that is the question.
In judging this problem we must introduce into our consideretions the question of whether the ordering system is used in a
bound index, in a file or in a multidimensional medium like
peek-a-boo cards, punched cards or a computer.

The conventional systems /bound indexes and files/ prefer
precoordination, and multidimensional records prefer postcoordination. In spite of this I think that the best way of judging
whether to precoordinate or postcoordinate is to ask: How will
the user seek the information? If he is likely to request a document about the "circulation of journals", he must be given
the possibility of searching under this precoordinated complex
term. If he is likely to request also documents about the "circulation" of other materials, it is necessary to enter this document also under this postcoordinated term. And if he is not
likely to demand the records about "journals" in general, we
shall not classify it under this term at all. This way of solving the problem also means a synthesis of two principles /precoordination and postcoordination/.

Last but not least come the two antitheses - alphabetical
and systematical arrangement of terms. Can we ever reconcile
these two principles? The answer is affirmative: moreover this
synthesis can be applied in two ways. In the first place by
using both an alphabetical index of terms and a systematical
display /tables/ in every ordering system.

However, within this, synthesis can be used even in the
ordering system itself. We have done so in our faceted classification in the INDORES information system. This faceted classification consists of four categories. The classes in the
categories are arranged alphabetically according to the capital
letters /which were chosen mnemotechnically/. The terms in the
classes are also expressed by mnemotechnical notation/in small
letters/ and again arranged alphabetically. The system is a
faceted classification with alphabetical arrangement of classes
and terms.

- 103 -

All   the   considerations   about
ordering   principles   apply   both
to   alphabetical   systems   and
classifications

ad 5. Interbranch relations

Thesauri are special ordering systems; this is their great
disadvantage in comparison with universal classification sys-
tems like D.C., U.D.C., Soviet Bibliographical-librarian Clas-
sification, Library of Congress classification etc. I doubt
whether a univeral thesaurus could be built for a simple reason
- when a thesaurus consists of more than 600 - 1,000 terms,both
the indexer and the user lose the survey of the thematical con-
tent of the information system. A larger system needs uncondi-
tionally thematical tables /or at least maps/.

Although one is aware of all the drawbacks of different
universal classifications, it is at least certain that they have
not the great disadvantages of all special thesauri - the problem
of marginal fields and the problem of auxiliary terms.

In every thesaurus the terms for marginal topics /chemistry,
physics, mathematics/ must be introduced again and again /always
in a different way/. There is no compatibility, although so many
thesauri need these terms. Evaluating the situation in Europe
now, one realises that the only coordinating factor in the ne-
twork of information centres - international and universal cla-
ssification - is disappearing and the network is disintegrating.
We are comming to an atomising stage - every branch and every na-
tion /in the same branches/is constructing its own thesaurus.This
is a problem of capital importance. On the other hand the intro-
duction of computerised systems demands deep analysis in index-
ing - and a special ordering system /whether it be the thesaurus
or a faceted classification/. How to solve this problem? It is
certainly one of the greatest problems of this and other confe-
rences on the problems of ordering systems.

My personal opinion is that we need a general classification
system forming a roof above the special ordering systems /both
thesauri and faceted classification in individual branches   of
human knowledge/. From this universal classification,   which

89

would not go into great detail /this would be the task of spe-
cial syste s/, we should use the term for all marginal fields
and all the auxiliary terms.

Who will work on the construction of this universal classi-
fication? In spite of the valuable work of the London Classifi-
cation Research Group it is clear that it would take years   to
construct a new system of universal faceted classification. And
reconstructing UDC for this purpose would also demand great
efforts. It is obvious that besides this, <u>it will be necessary
to state general principles for the construction of special or-
dering systems</u>.

BUILDING OF THE THESAURUS

David C. Weeks[x]

Two principal considerations in approaching the task of
thesaurus building are: the domain which is to be represented
and the sources from which concepts and terms are selected.

Whereas scientific and technical information systems have
tended to develop independently in the past, we must recognize
the inherent limitations of such a random pattern and seek
means of overcoming those limitations. Influences helping to
reverse this tendency stem from an awareness of the frequent
overlap between domains or even disciplines and from a growing
need for system-to-system communication.

The impact of these influences is apparent when we consi-
der the elements that comprise the language of a thesaurus de-
veloped for a specific application in a discipline. Note, how-
ever, that the discipline is most probably a subdiscipline of
some comprehensive discipline. This is largely a reflection of
the trend towards systems serving narrower fields but treating
them more intensively.

In these circumstances, a discipline-oriented thesaurus can
be expected to contain four vocabularies, each essential to the
literature of the domain and the retrieval interests of its
users:

- the core concepts and terms of the discipline - the langu-
  age essential to research and communication, describing the
  objects of concern and activities of the practitioner.
- the major discipline of which this domain forms a part -
  the scientific hierarchy illustrated by the domain of

[x] George Washington University BSCP, Washington.

entomology which is a sub-set of invertebrate zoology -
zoology-biology.

. other disciplines upon which the scientist or technician
draws in the performance of his work. These may be cross-
discipline domains such as chemistry/biology, physics/
chemistry; or they may be used as tools of research such
as statistics, logic, mathematics.

. the domains where this discipline may be applied - best
illustrated by the application of information science to
any scientific discipline.

Unless each of these elements is represented in sufficient
detail, the thesaurus cannot be effectively employed as a tool
for information control. It will fail to provide access to the
varied topics contained in a corpus of documents and it will
be unresponsive to the user's needs. As the interest in and
need for system-to-system communications becomes more intense,
the task of building a thesaurus becomes more complex. One way
of reducing the difficulty and at the same time promoting com-
patibility is the development of standard, uniform methods for
thesaurus building. This will encourage the symmetry and ba-
lan .hich are so essential to conversion from one system to
another and for one system to complement another. Another means
to the same end would be the development of standard components
which could be incorporated into several thesauri, since the
non-core elements are common to a number of domains.

Both possibilities require the combined efforts of spec-
ialists from various disciplines. The result of such projects
would greatly assist in halting the random development of
uncoordinated thesauri and of incompatible information systems.

S o u r c e s   o f   C o n c e p t s   a n d   T e r m s

The term thesaurus is becoming a common name for any infor-
mation system, indexing scheme or language. At the same time,
the concept of a thesaurus is also frequently equated with a
somewhat structured language in which broader, narrower and
related terms are arrayed - not always according to some for-
mula or rules that prescribe the level of terms. On the question

of methods of compiling a thesaurus, there are at least three
principles that must be considered in every instance, whatever
techniques may later be applied to the actual task of compila-
tion. These are: 1/ The source from which terms are to be se-
lected; 2/ The purpose for which the thesaurus is intended;
3/ The approach to construction.

Two conventional methods have been used for the selection
of terms. One method is the use of existing dictionaries and
other lexical aids. This approach, more likely to be used for
a discipline-oriented system, tends to produce a set of terms
that is representative of accepted usage. Terms are likely to
be relatively stable and to reflect a concensus of definition.
Most disciplines have a number of technical dictionaries as
well as encyclopedias, handbooks or similar basic resources
from which terminology can be derived. A disadvantage of this
method is that the vocabulary so assembled is somewhat "steri-
lized" and lacks the dynamic quality of current usage. Such a
thesaurus may be difficult to relate to the language of li-
terature or the patterns of recourse characteristic of poten-
tial users.

A second method - extraction of terms from a selected
sample of documents - uses primary rather than secondary so-
urces. Thesauri built from the language of literature are more
current, but risk the need for more frequent amendment as the
scope of literature is broadened in actual use. Terms of less
stability will be more frequent and defining notes more neces-
sary to specify the accepted meanings.Mission-oriented systems
are the more likely applications for theaauri constructed from
literature.

The purpose of the thesaurus is another issue that is es-
sential to its construction. Those intended for application
to the literature of a discipline are started with a different
set of premises from thesauri being developed for an organi-
zation with a defined mission and functions. While a core of
terms should be similar for a given discipline and a related
mission, considerable difference is likely in the peripheral
vocabulary. The two varieties are probably not interchangeable
in application.

A third consideration in methods of compilation is the
question of structure. In some instances, the structure is
built first and terms are selected and assigned at the time
of entry. Only in this way can a valid and workable thesaurus
be constructed. Reversing this process and building a struc-
ture after the terms are collected creates an uneven and
imbalanced structure not truly representative of the topical
universe it is meant to define.

SOME PRAXIOSEMIOTIC PROBLEMS OF SCIENTIFIC INFORMATION[x]

Tadeusz Wójcik[xx]

Praxiosemiotics /the theory of optimal sign/, praxiolinguis-
tics /the theory of optimal language/, the classification
theory, the theory of the designation of the message /the entry
characterizing the message/, the theory of automated informa-
tion retrieval - all these branches of science play an import-
ant role in the development of scientific information and its
theory.

In my report to the conference I present the principles of
my work on formal aspects of the general classification theory
and the theory of optimal language. I hope these considerations
will in some way stimulate the development of the central prob-
lem of scientific information, i.e. the theory of retrieval
languages.

---

[x] This report is based on T.Wójcik: Prakseosemiotyka. Zarys
teorii optymalnego znaku. /Praxiological Semiotics. An out-
line of a theory of the optimum sign/. Warszawa, 1969,
PWN pp. 289.
[xx] Warsaw University, Warsaw.

1. General description of the processes of information end cognition

I have subdivided cognition into sensory, symptomatic and sign-message as successive stages of cognition, terminating with the exclusively human stage.

The main problem is that of message which is a tool for converting an uninformed man, animal or machine into an informed one. Both message and any of its component-parts, of various degrees of complexity, are signs.

A message may be in the form of a sentence, or larger text, photograph, portrait, map, graph, a chair tipped against a table in a cafe to indicate the place is taken, a doorbell, etc.

A message may act in diverse forms and especially as sound and light waves.

The postulates of the optimum message require that it be best adjusted to three sides, namely:

1. to the addressee of the message,

2. to its sender or creator,

3. to the reality communicated.

These postulates are compared with the corresponding postulates of the optimum tool in general, which in this case is a message.

The postulates of the optimum message are the following:

1. concerning addressee of the message,

    1. maximum access /in space and time/,

    2. maximum perception of the form of message,

3. maximum clarity of message;

2. concerning the sender or creator of the message:

   1. maximum ease of learning the technology of making and
      sending the message,

   2. maximum facility in making messages /both their content
      and form/ and sending them;

3. concerning the reality which a message reflects:

   1. unambiguity of the message,

   2. maximum correspondence between the message and the
      reality it reflects,

   3. optimum exactitude of the message.

To a large degree, though not completely, the message de-
pends upon its components being optimum in order to achieve
an optimum effect.

The code of message M is the classification of the set
of components of message M. The postulates for the optimum
code are: maximum reconstructability of the message from, the
components of the code, maximum perception of the elements of
the code, and maximum ease of learning its elements.Generally
speaking, the optimum code of message M is that classifica-
tion of a set of elements that is necessary and sufficient
to make the message M optimum. This is the postulate of
isomorphism between a set of elements of the reality described
and a set of elements of a code that satisfy additional
postulates with regard to the addressee and the sender of the
message. A one-to-one correspondence is a particular example
of isomorphism.

2. C l a s s i f i c a t i o n   c o d e s

Classification codes are codes of classification messages
/i.e. classification/[x].

One of them will be described here since it seems necess-
ary for the understanding of the language code.

We make a table of classification assumptions to classify
the set S with respect to features A, B,...N. The columns of
the table will contain successive modifications of the features
A, B,..., N /see Fig. 1/.

P   — classificandum

		A	B	...	N
		1	2	-	n
0		A_0	B_0	...	N_0
1		A_1	B_1	...	N_1
2		A_2	B_2	...	N_2
...		...	...	...	...
		W_A	W_B	...	W_N

- criteria of classi-
fication/nonprecise/

- general forms of cri-
teris of classifica-
tion

- precised /detailed/
forms of criteria of
classification
/modification/

- criteria of classifi-
cation in precise form

Fig. 1. Table of Classification Assump-
tions /formal structure/

We may make different classification codes by transforming
the components of the table. It is sufficient to show in ad-
dition the substance of the classification /notation/. This may be
the codes of a "tree", a table, a group of words, letters or
digits, etc.

I have enclosed, in brief, a method of making an ordered
classification message in a positional digits notation which is

a particular example of the naming notation. The symbols consist of digits which are detailed numbers of lines including precise forms of criteria /modifications/. The position of a digit in the classification symbol indicates the number of the column which includes the corresponding modifications. Hence the positional notation. The resulting ordered classification is called systematization.

The components of a message /parts of a classification/are placed in a permanent position in relation to one another so that the classification message will be ordered.

Directives for making an ordered classification of the set P with respect to A /in detail: A_0, A_1,...,A_x/, B /in detail: B_0, B_1, ..., B_y/, ... N /in detail: N_0, N_1, ..., N_z/:

0. Make a table of classification assumptions of the set P with respect to the abovementioned detailed criteria /Fig. 1/.

1. Make an ordered Cartesian product of sets:

$$P \times /A_0, A_1, ..., A_x/ \times /B_0, B_1, ..., B_y/ \times .../N_0, N_1, ..., N_n/$$

which are mentioned in the table of classification assumptions.

2. Eliminate the elements of the Cartesian product where the digit 0 precedes every other non-0 digit /e.g. 011, 201/.

3. Insert the resulting symbols into columns containing the same number of non-0 symbols,

in increasing order, and keep the order of the Cartesian product.

By following the above directives in order, we get the following results in the case limited to 2 modifications, each with three characteristic criteria:

Directive 0

P	A	B	C
0	0	0	0
1	1	1	1
2	2	2	2

Fig. 2. Table of Classification Assumptions /digital notation/

O is the symbol for a classification criterium which is disregarded in a particular case.

Directives 1 and 2

```
P 000        P 100        P 200
          ┌─────────┐  ┌─────────┐
 P 001    │ P 101   │  │ P 201   │
          │ P 102   │  │ P 202   │
          └─────────┘  └─────────┘
 P 002
 P 010        P 110        P 210
 P 011        P 111        P 211
              P 112        P 212

 P 012
 P 020        P 120        P 220
 P 021        P 121        P 221
 P 022        P 122        P 222
```

Fig. 3. Cartesian Product. The eliminated
elements are framed

A set P of cloths from a textile factory will serve as an example.

The set is classified /systematized/ according to the following criteria:A: substance of cloth /natural, synthetic/, B: thickness of cloth /thin ≤ 0,1 mm, thick > 0,1 mm/, C: color of cloth /natural, dyed/.

Set table of ordered classification assumptions is the following:

Directive 3

One dimensional form	Two dimensional form Degree of systematisation			
	0	1	2 .	3
P 000	P 000			
P 100		P 100		
P 110			P 110	
P 111				P 111
P 112				P 112
P 120			P 120	
P 121				P 121
P 122				P 122
P 200		P 200		
P 210			P 210	
P 211				P 211
P 212				P 212
P 220			P 220	
P 221				P 221
P 222				P 222

Ordered classification — systematisation

Fig. 4. Systematization Message
in a Positional-digital Notation

Cloth P	Criteria of classification		
	A	B	C
	Substance	Thickness	Color
0			
1	natural	thin	undyed
2	synthetic	thick	dyed

Fig. 5. Table of Classification   Assumptions
/example/

The classification components of the different cloths are
as follows
P 000 = cloth /systematized set/,
P 100 = cloth of natural substance,
P 210 = thin cloth of synthetic substance,
P 121 = thick and undyed cloth of natural substance,
P 002 = dyed cloth.
     The table classification is a particularly useful form of
classification message. It has been widely applied to the classi-
fication-terminological standards in this country since the
numerical notation is brief.

3. L a n g u a g e   c o d e s

     Praxiological linguistics is a part of praxiological semi-
otica.The language is a classification of a set of sentential
message elements. The optimum language contains exclusively the
necessary and sufficient components for the optimum sentential
message to be in as aesthetic a form as possible.
     Two main tasks of praxiological linguistics are:
1. The task of cognition facilitates research on the basic lan-
    guage function on the basis of the simplest construction
  · /praxiological model of language/. Such a construction would
    also facilitate a description of natural languages which may
    be hypothetically regarded as a deviation from the model.
2. The practical task of direct utility in:
    1. research on problems of machine translation,
    2. research on the creation of the optimum language.

The optimum language, while being primarily the language
for scientific information, would enable direct communication
of anybody with anybody else. Moreover, it would enable the
formation of a better picture of the world, thus helping men to
function much more efficiently intellectually. The problems of
praxiological linguistics lie within the field of interest not
only of logicians and linguists but also praxiologists, cyber-
neticists, psychologists and sociologists.To get to the root
of the problem I have suggested an extensively developed group
of grammatical terms. The formal structure of the optimum lan-
guage is given priority in this elaboration which consists of
syntactic and phono-graphical structures. Next come the gener-
al principles of building the semantic structure of notion
groups, the content-structure. The last problem to solve is to
establish a one-to-one correspondence between the formal struc-
ture and the content.

Particular consideration should be devoted to the syntactic
structure of such sentential message components with increasing
degrees of complexity, es an element of a word /roots, affixes,
endings/, a word, phrase /a group of words like the subject and
its modifiers/, a simple sentence, a compound sentence.

The classification algorithm of the optimum language is a
set of instructions on how to make sentences or phrases at dif-
ferent degrees of complexity out of their lexical elements. The
algorithm also concerns the making of phonetic structure of the
language. Thus the algorithm is a formal instruction on how to
form language expressions. The classification method also touch-
es the problem of morphology - the problem of making correct
notion groups. Although the problems of the formal side of the
language can be easily solved, the content requires thorough
research. Such research should concern the entire set of no-
tion groups characteristic of different branches of science and
reflect these notions in the forms of the optimum language at
the formal, that is, syntactic and phonetic level. Thus the
algorithm can only concern the formal aspect of the language.

The optimum language model consists of the following parts:
1. the classification algorithm
   1. of the syntactic structure,
   2. of the phonetic structure,

2. morphemes,
3. examples of some semantic solutions,
4. the main principles of a one-to-one correspondence between
   the form and content of the language.

The classification algorithm of the optimum language con-
sists of a table of classification assumptions of the set of
sentential message components, and an instruction how to make
the ordered Cartesian product of the elements of the table..

Fig. 6 shows an outline of this algorithm.

C	Message components			
	A	B	...	N
0				
1				
...				
r				

Fig. 6. Table of Classification Assump-
tion /message/ /for table C/:
$\langle A \times B \times ... N \rangle$

W.: can get an algorithm of the syntactic or phonetic struc-
tures by substituting the variables accordingly.

The syntactic algorithm can be obtained in the form of one
table of assumptions. However, it would be very large and hardly
intelligible. Therefore, for the sake of convenience, we make
four tables for their respective components of different degrees
of complexity: one for the compound sentence; one for the simple
sentence; and one for the component of a phrase or word, re-
spectively.

Cs	Components of the sentence						
	Charac-teriser	Simple sentence	Conjunction	Simple sentence	Conjunction		Simple sentence
	A	B	C	D	E	—	N
0							
1							
...							
r							

Fig. 7. Table of Classification Assumptions /compound sentence/

Each of the four tables helps to obtain two results, namely:
1. A group of general atructures of a definite kind - symbolized by 0-1.
2. A group of language expressions of a definite kind - substituting an appropriate expression for the variables.
   The group contains 4 032 possible sentence structures.
   An algorithm of the phonetic make-up of a syllable is based on a table of claaaification assumptions like the one below.

		Description of components of a simple sentence													
		Subject—Object Phrase												Time-Space Indicator	
		Nominator 1				Relator				Nominator					
		Quantificatum				Quantificatum				Quantificatum					
S_2	Characterizer	Indicator	Attributor Degree		Quantifier	Indicator	Attributor Degree		Quantifier	Indicator	Attributor Degree		Quantifier	Time	Space
			1	2			1	2			1	2			
0	1	2	3		4	1'	2'	3'	4'	1	2	3	4	5	6
A	B	C	D		E.	F	G	H	I	K	L	M	N	O	P
0															
1															
r															

Fig.8. Table of Classification Assumptions /aimple sentence/

Ph	Indicator Attributor Quantifier	Word Conjunction	Indicator Attributor Quantifier	Word Conjunction		Indicator Attributor Quantifier
	A	B	C	D	...	N
0						
1						
...						
r						

Fig.9. Table of Classification Assumptions /phrase component-extended/

A syllable consisting of two initial and two final conson-
ants is regarded here as the most complex. There are the fol-
lowing types of syllables /here 0 stands for a vowel and 1,
for a consonant/:

$$
\begin{array}{ll}
00\text{--}1\text{--}00 & 0 \\
00\text{--}1\text{--}10 & 01 \\
00\text{--}1\text{--}11 & 011 \\[4pt]
01\text{--}1\text{--}00 & 10 \\
01\text{--}1\text{--}10 & 101 \\
01\text{--}1\text{--}11 & 1011 \\[4pt]
11\text{--}1\text{--}00 & 110 \\
11\text{--}1\text{--}10 & 1101 \\
11\text{--}1\text{--}11 & 11011
\end{array}
$$

	Word Components							
W	Semantic precisioner Sem I			Root (Rad)	Semantic precisioner Sem F			Syntactic pre-cisioner (Synt)
	Sem I₁	...	Sem Iₖ		Sem F₁	...	Sem Fₖ	
	A	...	E	F	G	...	M	N
0								
1								
...								
r								

Fig.10. Table of Classification Assumptions /word/

	Phrase				
0	1	2	3	4	5
1	1111	1111	1111	1	1
2	1101	1101	1101	0	0
3	1001	1001	1001		
4	1110	1110	1110		
	1100	1100	1100		
	1000	1000	1000		
			0000		

Fig.11.Group of structures of Simple Sentences

	Initial consonant			Vowel	Final consonant		
2	2	2	1		1	2	2
	A	B	C	D	E	F	G
0							
1							
...							
r							

Fig. 12. Sylleble

Such is the thorough description of the classification algorithm of the formal structure of the optimum language: the syntactic and the phonetic structure.

The grammatical morphemea are the following sets:
1. characterizer e.g.: it's so that, it'a not so that, let it be so, is it ao?
2. aentence conjuction e.g.; or, and if... then, if only,
3. quantifier e.g.: 0,1..., certain, every, few, many,
4. wordconjunction e.g.: or, and, nor, together with,
5. ayntactic precisioner e.g.: object, feature, relation, manner, place, time,
6. aemantic precisioner e.g.: -ness, -ry, -ism, -logy, etc.

I have given some specimens of lexical morphemes in this account.
Namely: the table of ayntactic precisioners and some semantic precisioners /relator precisioners, basic and auxiliary/, precisionera of gradat...n, relator precisioners of attributor /some spec'...ens/.

Besid 3, I have given the table of personal pronouns and a few quantifiers /numerals/.

Syntactic precisioner or syntactic type of word			Sentence function of word				
			Designating Designator		Attributing Attributor Quantifier	Function Conjunction	
						Word	Sentence
Semantic denomination scope of word		Object	Subject or predicate Nominator	Object marker -o	Attributor	Object -a		
	Feature	Described		Feature marker of an object -i		Feature -u	— •	-as
		Describing a state or relations: a kind of existence	Predicate Relator	Feature marker of type of existence -e	Adverbial	Attributor of type of existence -e		

*I introduce some pronounceable sound variables. Some precisioners are null (for a few lexical groups). The next examples are derived from Latin).

Fig. 13. Syntactic Precisioner

Semantic precisioner of relator (basic) Type of existence				Symbol	
				Number of phrases	
				1	2
1	General (to exist)			-ais	-ais
2	Functional existence	General (to function)		-ai	-ai-
3		Independent (o change into)		-er-	
4		Dependent	active (to change into)		-an-
5			passive (to be changed into)		-ai-
6	Extra-functional (to be a....)			-ao-	-ao-

Fig. 14. Semantic Precisioner of Relator /basic/

Examples (derivatives of „term")

termi	— temperature	buterman	— to cool
butermi	— cold	puterman	— to freeze
putermi	— frost	putermajno	— freezer
potermi	— cool	baterman	— to air-condition
batermi	— tepid	bitermajno	— radiator
bitermi	— hot	putermato	— frozen substance
pitermi	— heat	putermani	— freezing process
potermi	— warmth	etc.	

Absolute gradator			Comparative gradator		
Small	generally	bu	blu	<	Smaller than
	very small	pu	plu	the smallest	
	quite small	po	plo	<	
Medium		bo	blo	Equal	
Large	generally	bi	bli	>	Larger than
	quite large	po	plo	>	
	very large	pi	pli	the largest	

Fig. 15. Semantic Precisioner of Feature Gradation /gradator/

Some examples of the use of the relational precisioner of an attributor:

1. how existing, functioning? /null precisioner/
   domo ligna - wooden house, scriban bone - to write correctly
2. related to /whom/? re-
   re-Petra frato - Peter's brother
3. Whose property? di-
   di-Petra domo - Peter's house
4. making, doing what? la-
   la-libra scribano - /a man/-writing a book
5. made, done by whom? par-
   -par-Petra libro - a book written by Peter.
6. coming from where? de-
   de-Africa frukto - an African fruit /from Africa/
7. destined /suitable for/? pur-
   pur-homela vesto - women's wear /for women/

So much for the syntactic structure of language. One can especially extend other semantic precisioners. The phonetic structure of language requires experimental research by phoneticians and psychologists.

This means choosing the optimum sounds and the syllable make-up and then the prosodic features of the language.

The application of the optimal language to the needs of scientific information opens broad perspectives for improving the work of information systems.

108

ixtor and syntactic derivatives

Fig. 16. Table of Word Classification Assumptions

2. Another word

General subject/object (personal pronoun)				Generally	Particularly		
					Gender		Neuter
					masculine	feminine	
„Person"	0	someone something	Plural/Singular	generally	lo	ilo	elo	olo
				singular	prolo	prilo	prelo	prolo
				plural	solo	silo	selo	solo
	1	I we		generally	mo	imo	emo	
				singular	premo	primo	premo	
				plural	semo	simo	semo	
	2	you (sg.) you (pl.)		generally	vo	ivo	evo	
				singular	provo	privo	prevo	
				plural	sevo	sivo	sevo	
	3	he(she), it, they		generally	ro	iro	ero	oro
				singular	prero	priro	prero	prero
				plural	sero	siro	sero	soro

pre—one; se—more than one

Fig. 17. Personal Pronoun

Terms of existence for root <ved>				Syntactic type					
				Nominator			Attributor of		
				Existor	Feature marker	Object marker	Existence	Feature marker	Object marker
				–	-d	-o	-o	-u	-i
Existence	functional		generally	– (sh)		informative cognition process	element of informatively cognitive process			
			general	ed	to function informatively	informative functioning		informatively		informative
		independent		er	to get to know	cognition process	the one who gets to know	cognitively		cognitive
		dependent	active	en	to get informed	informing process	the one who informs	informingly		informing
			passive	ef	to be informed	the state of being informed	the one who is informed			knowing
	extra functionally			es	to know	knowledge	the one who knows	with knowledge		

Fig. 18. Family of words /example/

ON A DEFINITION OF A THESAURUS SYSTEM AND THESAURUS STRUCTURES

Irena Bellert and Olgierd A. Wojtasiewicz[x]

In the papers submitted for the International Conference on
General Principles of Thesauri Building, held in Warsaw in March
1970, and in the discussions which took place during the Confe-
rence, the terms "system of thesaurus", "structure of thesaurus"
have been used repeatedly in several contexts without being pre-
cisely defined - moreover, in different senses of these terms.
We believe that a formal definition of the system of thesaurus
that would account for the relations involved would make it poss-
ible not only to use these terms consistently, but also would be
helpful in constructing thesauri which pertain to selected areas
of science or technology in accordance to the same well defined
scheme, independently of the language in which the given the-
saurus is described. A comparison of thesauri in two different
languages could thus be arrived at more easily and multilanguage
thesauri could more conveniently be constructed.

The present proposal can be treated as a comment and an out-
line of a formal description of the ideas contained in the UNESCO
Guidelines for the Establishment and Development of Monolingual
Scientific and Technical Thesauri for Information Retrieval, a pa-
per submitted for the International Warsaw Conference on General
Principles of Thesauri Building. It goes without saying that for
a more specific formulation of the relations discussed below, it
would be necessary to elaborate on this topic in connection with
a given specific area of science or technology.

_____

[x] Department of Formal Linguistics, Warsaw University

Any system can be defined as an ordered sequence

$$\langle A, R_1, \ldots, R_n \rangle$$

where A is a set and R_i /for $1 \leqslant i \leqslant n$/ is a relation
defined on A.

A thesaurus can be defined as a specific system:

/1/ $$\langle T, R_1, \ldots, R_n \rangle$$

where T is a finite set of terms /nouns and noun phrases of
a natural language/, and the relations $R_1, \ldots, R_n$ are one- or
two - place relations defined on T. Certain relations will
be characteristic of a specific domain of science or techno-
logy, others will be characteristic of any domain for which a
thesaurus may be constructed. Among the latter relations we
may mention a linear order relation corresponding to the
alphabetical order; partial order relations corresponding to
concepts such as "term broader than", "term generic to" /and
their respective converses: "term narrower than", "term speci-
fic to"/, "part of whole" etc. equivalence relation, such as
synonimity relation between terms; other relations correspon-
ding to concepts such as "used for", "belongs to the domain
of", "thing-property", etc.

It is worth noting that every relation determines a graph,
which would correspond precisely to what is referred to as
"graphical display" in the description of thesauri. The term
"structure" as used in connection with thesauri can thus be
identified with the structures of the graphs determined by the
corresponding relations. In this connection the terms "system"
and "structure" as used in the description of thesauri would
be interrelated and precisely defined. For we have a set of
structures given by each thesaurus described as an ordered
system consisting of a finite set of terms and a sequence of
relations.

Let us now present some examples of relations which can
be defined for any thesaurus, and give some additional com-
ments.

The synonimity relation R_{syn} is an equivalence relation defined on T, for it is:

/a/ reflexive R_{syn} /x,x/

/2/ /b/ symmetric R_{syn} /x,y/ $\Rightarrow$ R_{syn} /y,x/

/c/ transitive: R_{syn} /x,y/ $\wedge$ R_{syn} /y,z/ $\Rightarrow$ R_{syn} /x,z/

R_{syn} - being an equivalence relation - forms a partition on the set T, that is, it divides the set T into non-empty, disjoint subsets /equivalence classes/. We may then choose a representative element x_o from each subset that contains more than one element, and the term chosen as a representative, will then be used in the description of other relations, all the other elements will appear only in the relation of alphabetical order, and will not be used anywhere else.

The set T will comprise all the possible terms,descriptors and non-descriptors. The set T can thus conceived   of as

/3/ $T = D \cup N$

to be read thus: the set T of terms is the union /logical sum/ of a set D of descriptors and a set N of non-descriptors;

   D and N are finite, non-empty disjoint sets:
   $D \cap N = \emptyset$
   The following relations can also be defined

/4/   P /a,b/

/5/   B /a,b/

/6/   R /a,b/

P/s,b/ is the preference relation, and may be interpreted thus: $\underline{a}$ is to be preferred to $\underline{b}$ as a descriptor. This relation is obviously irreflexive, asymmetric, and intransitive. If a term is once rejected as unsuitable to be a descriptor, then it cannot on any other occasion be given preference to any other term, because if it were given preference, then it would have to be included in the list of descriptors in spite of being previously disqualified as a descriptor.

B/a,b/ is the relation holding between a broader and a narrower term: a is a broader term than b. This relation is irreflexive, asymmetric, and transitive, and as such it is an ordering relation. The relation holding between a narrower and a broader term, N/a,b/, also listed in the Guidelines, can trivially be defined as the converse of the former:

/7/ $$N/a,b/ \Longleftrightarrow \breve{B}/a,b/ \Longleftrightarrow B/b,a/.$$

R/a,b/ is the affinity relation, and is reflexive and asymmetric, but not transitive: although transitivity may be observed in certain cases, it may not be assumed to be a rule. This can be explained by the following example: steam and steam engine are related terms, and so are steam engine and combustion engine, but steam and combustion engine presumably are not related.

The following implications are assumed:

$$P/a,b/ \rightarrow //a \in D/ \wedge /b \in N//,$$

/8/ $$P/a,b/ \Rightarrow R/a,b/,$$

$$B/a,b/ \Rightarrow R/a,b/.$$

The following relation can be defined by the well-known procedure:

/9/ $$B'/a,b/ = B/a,b/ \wedge \neg \bigvee_c /B/a,c/ \wedge B/c,b//.$$

This is to be read thus: a is the next broader term to b /a is broader than b and there is no term c such that it stands between a and b in the hierarchy of broadness/.

In the definitions below it is assumed that a and b are in D, and hence this fact is not marked in the formulae.

/10/ $$A/a/ = \left\{ b: R/a,b/ \right\}.$$

This is the affinity class of a.

/11/ $$P/a/ = \left\{ a \right\} \cup \left\{ b: B/a,b/ \right\}.$$

114

The field of a: a and all those terms to which a is broader. In the definions the one-element set $\{a\}$ must be included since B is not reflexive, B/a,a/ accordingly does not hold, and thus a would not be in its own field. This does not apply to /8/ as R is reflexive.

/12/ $H/a/ = \{a\} \cup b:\{/B/a,b/ \vee B/b,a//\}$.

This is the hierarchy class of a, which, as compared with F/9a/includes not only those terms which are narrower than a, but also those which are broader than it.

Usually in a hierarchy class limits must be set beyond which certain terms are not included because of being either too general or too specific. This can be defined as follows:

/13/ a = minimally narrow term/s/ in H/a/⟺

$$\bigvee_{\Longleftrightarrow b \in H/a/} /B'/a,b/ \wedge /a \in D/ \wedge \neg /b \in D//,$$

/14/ a = maximally broad term/s/ in H/a/⟺

$$\bigvee_{\Longleftrightarrow b \in H/a/} /B'/b,a/ \wedge /a \in D/ \wedge \neg /b \in D//.$$

It is to be noted that the preference relation does not ensure the inclusion in D of all those terms which it may be desirable to have as descriptors. This is so because this relation provides for the inclusion in D of those terms which are preferred to some other terms. It may happen,however, that a term is to be included even though its inclusion does not imply the exclusion from D of any term. If, for instance,one makes a thesaurus of chemical terms, he will include as descriptors the names of all chemical elements. Hence, if oxygen is included, this fact does not bar any other term from inclusion in D under the preference relation. This is why it is recommended that an inclusion operation be introduced,to be defined thus:

/5/ $I/a/ \Longrightarrow /a \in D/$.

Obviously,

/16/ $P/a,b/ \implies I/a/$,

but not conversely.

The Guidelines mention the whole-part relation as a case
of the relation holding between a broader and a narrower term.
It is to be noted, however, that B holds between terms, and
the whole-part relation holds between designata of terms /that
relation can also hold between terms in the sense that _combus-
tion_ is part of _combustion engine_, but this obviously is not
what the authors of the Guidelines mean/. If this relation is
to be retained, the following improvement is suggested. A re-
lation M holding between terms is introduced and is tentati-
vely defined thus:

/17/ $M/a,b/ \iff$ there exist x and y such that x and y
                are objects, and x is part of y, and a and
                b are names of x and y, respectively.

The concepts defined or introduced above can be used in a
further extension of the formal description of the system under
consideration. E.g.,

/18/ $F/a/ \subset H/a/$,

/19/ The fields of $\underline{a}$ and $\underline{b}$ overlap $\iff F/a/ \cap F/b/ \neq \emptyset$.

As the main goal of the construction of thesauri is their
use in retrieval systems, it would be desirable to describe a
thesaurus system as a component of a retrieval language. An ana-
logy suggests itself with a description of a natural language
system. The search for a recursive system of rules for genera-
ting all and only sentences of a natural language has been in
the centre of attempts made by a number of recent linguistic
projects. It has appeared that the problems of grammar and
lexicon are deeply interconnected, and no adequate description
of lexical entries can be achieved without syntactical infor-
mation. It seems that, analogically, there is an interconnec-
tion between a thesaurus system and a retrieval language. The
retrieval language to be used for a thesaurus should be cons-

- 131 -

tructed in close relationship with the structures determined by
the relations defined on the set of terms T, as discussed above.
It is necessary to specify what sort of questions can be asked in
a retrieval language, and this cannot be done otherwise than by
relating them to ''e thesaurus system which has to be precisely
defined for that purpose. It may be worth mentioning here that
those who work on constructing thesauri and retrieval systems are
in a better position than linguists, for they can establish nor-
mative, rather than descriptive rules/they can get rid of syno-
nyms, homonyms or vague terms/.

In conclusion, we wish to emphasize that the problem of a for-
mal description of thesauri has been only touched upon inconclu-
sively, but the matter seems worth elaborating on, and our goal
was to draw the reader's attention to such a possibility.

MINUTES OF THE CONFERENCE

Monday, 23rd March 1970. Afternoon Session. Chairman: J. Toman

POINT 1 OF THE QUESTIONNAIRE: DEFINITION OF A THESAURUS

1.1 What is meant by a thesaurus?

Mr. T. M. A i t c h i s o n: "A thesaurus is an alphabeti-
cal listing of concepts /i.e. descriptors/ which provides struc-
tural and relational information about the concepts".

Mr. J a n s e n: For purposes of information storage and
retrieval a thesaurus is an orderly compilation of concepts
    - represented by as many synonymous terms as possible in one
      or more languages,
    - in which homonymous terms are specially marked,
    - in which a descriptor univocally represents a concept, and
    - in which semantic relationships between concepts are regi-
      stered.
Underlying definitions:
    1/ Definition of Concept:
    Mental idea of material or immaterial object based on common
characteristics which are usually formed by abstraction and found
identical.
    2/ Definition of Term:
    Name given to a concept and consisting of one or more words.
    3/ Definition of Descriptc :
    Univocal representative of a concept in a documentation sys-
tem. The descriptor can be a fixed term /"preferred term"/ or
any other stipulated designation.

Mr. L e s k i: "A thesaurus may be defined as follows: An open system covering a determined full thematic range containing an orderly multitude of terms, some of which are admitted as descriptors, showing the relations between these terms and their mutual dependence".

Mr. R o l l i n g: "A thesaurus can be defined as a structured vocabulary for use in information storage and retrieval systems".

Mr. W y s o c k i cited the definition from the UNESCO Guidelines: "By the word thesaurus is meant a comprehensive and structured vocabulary of interrelated terms some of which are used in the indexing and retrieval of a collection of documentary material pertaining to a specific domain or domains of science and technology".

Mr. T o m a n pointed out that a definition of an object must distinguish it from other similar objects but unfortunately in most of the cited definitions the words "classification", "uniterms" or "subject headings" can be inserted instead of the expression "thesaurus", and still the definition would keep its sense.

The session decided to postpone this question to Wednesday or Thursday when other problems of thesauri would be discussed.

Mr. T o m a n proposed to build the definition /or better, explanation/ by expressing first the purpose of the thesaurus and then by describing its possible characteristics: "A thesaurus is a system of controlled terms used for characterizing the content of documents in a storage and or retrieval system. Conventional thesauri use alphabetical listing for displaying synonyms, hierarchy, and other relations, whereas thessuri with graphic maps stress the importance of the systematic display of terms. Thesauri are further characterized by their dynamics, by a weak hierarchy, by preferring post-coordination to pre-coordination. In contrast with classification systems they show related terms other than synonyms, broader and narrower terms, but do not use notation or categories".

Mr. V a r g a considered some requirements regarding the terms of a thesaurus, pointing out that they must

- be based on a natural language,
- be unambiguous,
- constitute a decentralized collection.

As to relations within a thesaurus, they must
- express only objective, not artificial connections between terms,
- show many aspects of a connection
- be clearly distinguished
- be reciprocal

1.2 Structural elements

The participants in the session indicated several elements characteristic of thesauri:
- dynamics of the thesauri /a better word than "open-endedness"/;
- display of hierarchical, synonym and associative relations;
- display of relations: term to term, concept to concept, concept to term.

1.3 Factors influencing the organization of a thesaurus:

Messrs. Lloyd, Gravesteijn, Leski, Molnár, Mojžišek, and Maixner stated these factors:
- branch of science or technology
- thematical range and overlapping
- assumed degree of fineness
- users' requirements
- grammatical rules
- structure of the file
- methodology of the system
- clearness or vagueness of the terminology of the branch
- language /Hungarian or German/
- nature of the medium of the record /whether a bound index, a card file or a computer/.

Mr Rolling completed this list by citing his formula for the size of a thesaurus:

$$T = a\,k\ \sqrt[n]{\frac{V}{R\ k}}$$

T - size of thesaurus
a - indexing depth
k - redundancy factor /number of synonyms/
n - retrieval strategy /numbers of terms according to the strategy/
V - size of document collection
R - response /number of references expected by the type of user/

With regard to the structure of thesauri Mr. V a r g a stated that thesauri must be:
- compatible
- easily broadened
- easily corrected

Tuesday 24th March, Morning Session. Chairman: Mr. L.Rolling

ITEM 1 ON AGENDA: THE ROLE OF THESAURI

The main use of a thesaurus is for terminology control in
indexing and retrieval.
Agreement was reached on the following:
The role of a thesaurus is to ensure that indexing and retr-
ieval of documents can be effected with maximum efficiency /pre-
cision and completeness/.
In addition, the existence of a thesaurus covering a subject
field tends to stabilise the terminology of this field.

ITEM 2b ON AGENDA: THE DEFINITION OF A DESCRIPTOR

The Chairman analysed the definitions proposed by Messrs.
L e s k i,   J a n s e n,   S p a n g - H a n s s e n   and
M o l n á r, and the definition contained in the UNESCO Draft
Guidelines. They agreed that:
   - a descriptor is a formalised, standardized, or controlled
     term;
   - a descriptor is to represent one /or a combination of/ con-
     cepts in an unambiguous, or univocal way;
   - and that descriptors can consist of symbols.

Mr. J a n s e n   thought that notations must also be con-
sidered as descriptors; Mr. R o l l i n g   was of the opinion
that notations such as classification codes or current   numbers
assigned to descriptors for further processing cannot be conside-
red as descriptors.

Messrs. R o s e n b a u m   and   P o l e t y ł o   wished
to limit the use of descriptors to systems based on concept coor-
dination; subject headings should not be considered as descrip-
tors.
The assembly appeared to agree on the following:
A descriptor is an authorised and formalised term or symbol
in a thesaurus used unambiguously to represent the concepts of
documents and queries in information systems based on concept
coordination.

ITEM 2c ON AGENDA: THE DEFINITION AND NAME OF FORBIDDEN TERMS

The assembly preferred the name NON-DESCRIPTOR to the pro-
posed name ASCRIPTOR.

By definition, any thesaurus term or symbol not considered
as a descriptor is a non-descriptor.

ITEM 2d ON AGENDA: REQUIREMENTS TO BE FULFILLED BY A DESCRIPTOR

It was agreed that a descriptor must unambiguously charact-
erize the concept/s/ that it represents, and that its spelling
should be subject to a number of rules.

Agreement was also reached on the requirement that a descrip-
tor must have a "reasonable" frequency of assignment and possess
a certain combinatory power. A "reasonable" frequency could be
defined, according to Mr. Molnár, as a function of aver-
age frequency of assignment.

Tuesday, 24th March 1970. Afternoon Session. Chairman: J.Lloyd

The meeting reconvened at 3.00 p.m. to continue its consi-
deration of point 2 of the revised agenda: "What should be de-
manded in order to accept a term as a descriptor". The question
of grammatical form was placed before the group and Mr. W y-
s o c k i read from page 5 of the UNESCO Guidelines onward.
Mr. S p a n g - H a n s s e n observed that the proposed
UNESCO rules concerning "number", "noun f __", etc. were not
applicable to many inflected European languages.

Mr. W y s o c k i pointed out that the Guidelines were
written in English, for English, as stated in the document.

General opinion was that since grammatical usage varies
from language to language, it was better for each thesaurus to
establish its rules, state what they were in a preface, and stay
absolutely consistent with them throughout.

The meeting then moved to point 2e of the revised agenda:
"The terminology of cross-references".

Mr. V a r g a questioned the Guidelines approach to cross-
references between various thesauri.

Mr. W y s o c k i explained that the purpose of the Guide-
lines was not directed to the construction of any one thesaurus.

Mr. V a r g a then spoke on relations within a thesaurus
and the significance of hierarchical and non-hierarchical con-
cepts both within and beyond a thesaurus.

Mr. S p a n g - H a n s a e n asked if we were discussing
"roles" and if so, did they belong in a thesaurus?

The C h a i r m a n directed the discussion back to "pre-
ferential, hierarchical, and affinitive" as proposed in the
Guidelines.

Mr. R o l l i n g suggested that concepts might be linked
in an A + B relationship by "generic posting" as discussed in
his paper. He preferred this procedure to posing a Boolean query
in retrieval.

Mr. T o m a n suggested that aa terms can be members of
a plurality of hierarchical chains, all relationships should
be shown by tables or a thematic display.

Mr. J a n s e n agreed and offered a citation from E.Wüs-
ter: "Die Struktur der aprachlichen Begriffswelt und ihre Dar-
stellung in Wörterbüchen".

The aasembly agreed to accept aection X of the UNESCO
Guidelines.

It was agreed to postpone point 3 of the revised agenda
until after formulation of a definition, and discussion was
continued on point 4 a: "The methods of building thesauri".

. Mr. T o m a n proposed two thesauri building methods:
from the bottom and from the top. The first would involve col-
lecting terms at random from experts, other thesauri,glosaaries,
etc. - then deciding on hierarchy, synonyms and homonyms."From
the top" implied selecting large classes, subdividing them more
specifically, and continuing in this fashion.

Mr. M o j ž i š e k briefly explained the statistical
approach being developed at his institute in Prague and referred
the group to his printed report on the subject.

Mr. L e a k i explained, with schematic diagrams, the sy-
atem being used in Warsaw, which combines a theoretical and prac-
tical approach.

Mr. S c h i f f explained the KWIC indexing being used at
the Central Technical Library in Budapest in order to provide
subject specialists with information as quickly as possible. Ex-
perts then underscore the relevant words from the KWIC for each
document and return the information to the library. They are thus
building up a file of subjectively-approved descriptors that will
be used in the future for thesauri building.

Discussion turned to "overlapping fields", and Mr. L e s k i
said that a thesaurus cannot be narrower than a discipline. He
graphically displayed the problem of discipline overlap between
chemistry, physics, and biology. He stressed that hierarchies
may change within the area of overlap.

Mr. M o j ž i š e k stated that his statistical method
would solve overlap.

Mr. G r a v e s t e i j n   discussed the difficulty oi
resolving overlap problems with disparate disciplines, or eveu
sub-disciplines such as exist in geology.

Mr. S p a n g - H a n s s e n. Concerning singular and
plural, cf. UNESCO Guidelines, VIIc:

"The use of singular versus plural form to denote different
concepts /cf. the example WOOD versus WOODS in the UNESCO Gu-
idelines/ is comparable to the use of qualifiers for homonyms
/cf. the example BEAMS /ELECTROMAGNETIC/ versus BEAMS /STRUCTU-
RAL/, and since there are languages in which a plural ending is
not always present /e.g. Swedish/, the use of qualifiers seems
recommendable as a universal tool".

"In cases where only the singular or the plural form of a
given word is included in a thesaurus, the question is reduced
to a pure matter of expression, and it seems recommendable that
either singular or plural can be used consistently".

Concerning the reference aystems suggested by Mr. V a r g a
and Mr. J a n s e n:

"The more eleborate systems of relationships, including as-
pects that have been suggested by Mr. V a r g a and Mr. J a n-
s e n   eeem to me comparable to the well-known roles introdu-
ced by Costello and others as a syntactic supplement to Taube's
uniterms.

"However, it is a new point of view to include indicators
of aspects etc. in the thesaurus itself, which is usually re-
garded as a semantic /not a syntactic/ tool, comparable to s
vocabulary".

Wednesday, 25th March 1970, Morning session.
Chairman: J. Gravesteijn

After approval of the minutes of the meetings on March 24th,
the discussion continued on point 4a: methods of building thes-
auri for overlapping fields.

Mr. T o m a n   mentioned the fact that thesauri are not the
first tools of indexing and retrieval in the history of documen-
tation and that many publications have been written on subject
headings, conventional classification systems and faceted class-
ification in the past. This knowledge and experience should be
used for the benefit of the construction of modern thesauri.

In conventional information systems UDC played an important
role in many countries, being the integrating factor in the net-
work of information centres. Introducing special thesauri would
mean atomizing this network. Each information centre in each
country is building its own thesaurus without regard to neighbo-
uring fields and to similar information centres abroad.

On the other hand it is clear that a mechanized information
system needs deep indexing and this demands an ordering system
made to measure and not a universal system. We must realize that
the problem of the relation of special and universal ordering
systems is of capital importance to this and other similar con-
ferences. The question of relations between the thesaurus for a
special field and some superstructure, perhaps of very generic
character, must be solved.

Mr. R o l l i n g   gave an example of compatibility between
several thesauri in the same field /metallurgy/ and of compatibi-
lity between thesauri in different overlapping fields /nuclear
sciences and metallurgy/.

UNESCO is fully aware of the problem of proliferation and
Mr. W y s o c k i mentioned the existence of the two clearing
houses in Warsaw and Cleveland dealing with thesauri.

Mr. L a s k i   stated that there must be valid reasons which
impelled information specialists to work on other ordering sys-
tems than classification systems. After having changed from a
formal to a language system, the problem is now to establish com-
patibility between the existing thesauri.

The disadvantages of classification systems were discussed
by Mr. Mojžíše', Mr. Jansen, Mr. Aitchison and Mr. Lloyd, as
far as file handling in overlapping fields and communication
between user and file are concerned.

Mr Wysocki pointed out that an integration of the-
sauri in a general system is one of the aims of the UNISIST
project. In this project two systems are distinguished:

a/ A broader scheme of classification

b/ Thesauri for the different branches of knowledge.

After discussion the assembly concluded on the proposal of
Mr. Rolling that we can distinguish three different levels in
ordering systems:

1/ A specific level serving for indexing and retrieval pur-
poses.

2/ A general level to assure compatibility between thesauri.

3/ A classification system that plays an organizing role.

The factors influencing the compatibility of thesauri in
related fields are:

- homonyms
- synonyms
- relations between terms belonging to different fields
- point of view when considering descriptors from one field
  or another.

The need for cooperation in establishing thesauri in the same
field and in overlapping fields was admitted by the assembly.

Mr. M o j ž i š e k stated, however, that the problems re-
lated to overlapping fields can only be resolved from the point
of view of each field concerned.

A lecture by Mrs. B e ļ l e r t on the linguistic approach
to thesauri building was a welcome contribution to the discuss-
ion, and it clarified some of the terms used earlier.

Point 4a: Methodological problems in the establishment of mul-
tilingual thesauri.

The chairman described the activity of ICSU-AB. The project
for a multilingual thesaurus in the field of geology was mentioned
/English, French, German, Russian/.

Some complementary information on existing multilingual the-
sauri was given by Mr. R o l l i n g. The languages used in

the case of the thesaurus of the DIRR /Roads/ are English,
French, and German. The languages are English, German and Ita-
lian in the case of the thesaurus of the "Centro Sperimentale
Metallurgico". The European Committee is also working on a mul-
tilingual thesaurus.

Mr. L e s k i   reported that a multilingual thesaurus in
the "Science of Science" field for Polish, German, Czech   and
Russian has been built.

Mr. M a l m s t e n, who had already drawn attention to
semantic problems when comparing thesauri in different languages,
mentioned the work done by the "Comité international des arts et
des traditions populaires" /8 languages/.

  According to Mr. Spang-Hanssen, experience with multilingual
thesauri in Scandinavia showed that these thesauri can only be
established in fields with fixed terminology.

Wednesday,25th March. Afternoon Session. Chairman: T.M.Aitchison

---

POINT 4a ON AGENDA: METHODS OF BUILDING THESAURI

The discussion of this item was continued from the morning
session.

Mr. J a n s e n   stated that the IDC thesaurus is multilin-
gual in German and English and to a lesser extent in French. It
is used by them in indexing papers in these languages.

Mr. L e s k i   explained that his group used only descrip-
tors and candidate descriptors. If a precise descriptor could
not be found an auxiliary term in a foreign language might be
used, but controlled terms were required in social sciences.

Mr. R o l l i n g   stated that there were two types of mul-
tilingual thesauri: those which were built up simultaneously in
each language, and those in which the thesaurus in one language
was translated into other languages.

Mr. R o b o w s k i   referred to his experirce in diction-
ary preparation. It was impossible to translate terms from one
thesaurus to another: instead one should translate concepts, i.e.
descriptors as classes, not according to the exact meaning of
the words.

Mr. M a i x n e r   considered it advisable to translate
whole sentences rather than word for word. One might try to do
this by machine, but the problem of machine translation had not
been solved.

Mr. L l o y d   did not agree that one could not translate
one thesaurus into another.

Mr. J a n s e n suggested that while 90% of the words might
be translated satisfactorily, in 5% to 10% the translation would
produce noise.

Mr. L e s k i   agreed with Mr Rolling's division of multi-
lingual thesauri into two types. He considered that the same
scheme must be used for both thesauri if they were to be compa-
tible.

Mr. V a r g a   suggested that concepts were different in
different languages and that one could translate from a language
with broad concepts, but not the other way round.

- 146 -

Mr. R o s e n b a u m   suggested that concepts were mental
things, but were expressed in words.

Mr. S p a n g - H a n s s e n   considered that the confe-
rence was confusing translation of items of information with
translation of vocabularies.

Mr. L l o y d   stated that Mr Jansen's 5% was dealt with
in other situations by retaining the word in the form of the
other language, i.e. by not translating it.

Mr. M o j ž i š e k   reminded the conference that it was
concerned with technical information, and that definition and
precision were required.

Mr. M o l n á r   considered that the conference was not ob-
liged to reach agreement on methods, and that the problems bel-
onged to the science specialist rather than the information
specialist.

Mr. J a n s e n   explained that although one can translate
any sentence from English into German or vice versa, the sub-
divisions of the English term might be very different from those
of the equivalent German term.

Mr. L l o y d   thought that if it were possible to trans-
late thesauri in special subject fields, it was possible to build
thesauri in different languages within these subject fields.

Mr. R o l l i n g   stated that the assembly could either
go on with the discussion for some days or reserve a time at this
conference or later, depending on when the UNESCO Guidelines
dealing with multilingual thesauri would be ready.

Mr. W y s o c k i   said that the first draft was planned
for July 1970.

It was agreed to leave the discussion of multilingual the-
sauri and to move on to the next item on the agenda.

POINT 4b ON AGENDA: METHODS OF BUILDING DESCRIPTORS

Mr. W y s o c k i   drew attention to Items VI: "Selection
of descriptors" and IX: "Methods of entering" of the Guidelines.

Mr. R o s e n b a u m   considered that the problem had
already been discussed and suggested that the meeting proceed
with other items.

131

Mr. M e l m s t e n   suggested that in building descrip-
tors one should be concerned elso with the topology or struc-
ture /neighbourhood/.

Mr. R o l l i n g   mentioned thet descriptors were added
and eliminated throughout the life of a thesaurus end pointed
out that this was included in the Guidelines.

Mr. T o m e n   referred to items other than descriptors
or nondescriptors: cendidate descriptora and explicative words
added to descriptors. He also mentioned descriptors used in
parentheses in abstracts.

Mr. A i t c h i s o n   mentioned the problem of degree of
pre-coordination.

Mr. T o m a n   steted that this depended on the require-
ments of the user.

Mr. R o s e n b e u m   considered that machines allow eny
level of pre-coordination without any difficulties.

Mr. J a n s e n   recommended that compound concepts should
not be split up into less than single independent and unambig-
uous concepts.

The delegates agreed to provide UNESCO separately with any
comments on the relevant sections of the Guidelines they wished
to make.

POINT 5 ON AGENDA: CONDITIONS WHICH DESCRIPTORS AND THESAURI
MUST FULFIL IN ORDER TO ENSURE THEIR INTER-BRANCH AND INTER-
LANGUAGE COMPATIBILITY

It wes agreed that these points had already been diacussed.

POINT 6 ON AGENDA: CONDITIONS WHICH MUST BE FULFILLED BY DES-
CRIPTORS AND THESAURI AS TOOLS FOR FURTHER DEVELOPMENT OF IN-
FORMATION

Mr. L e s k i   explained that it might be that the thesaurus
would only be suitable for uae with given aims and at a given
period of time and might not be adaptable for other information
processing conditions. On the other hand, it may be a tool which
could be further developed in roles other than retrieval, for

example in machine analysis of texts, or in the synthesis and
analysis of information.

Mr. R o s e n b a u m   said that thesauri might be used
for other purposes, including developing a science and finding
relations between concepts.

Mr. M a l m s t e n   suggested another use: in standard-
ising glossaries where a thesaurus rather than a list of terms
would be provided.

Mr. R o b o w s k i   doubted the future of thesauri since
they could not be built up by machine and were not necessary
for the machine manipulation of large numbers of documents.

Mr. W e e k s   suggested that thesauri might be suitable
for 1960 but that something different would be required in the
next few years.

Mr. M a l m s t e n   pointed out that a number of impor-
tant information service required no thesaurus and suggested
that this point should be discussed.

Mr. L l o y d   considered, however, that at the present
stage of the computer art, with comparatively small stores, high
input cost, etc. thesauri were necessary.

Mr. R o b o w s k i   suggested the use of glossaries or
lists of keywords in fields with specialized vocabularies and
for small collections of documents.

Thursday, 26th March 1970. Afternoon session.
Chairman: H.Spang-Hanssen

The conference turned to point 7 of the revised conference
order, viz. organizational problems; subquestions /a/ and /b/
were treated together.

Mr. W y s o c k i reported on UNESCO's activities: guide-
lines for monolingual thesauri, to be followed during 1970 by
guidelines for multilingual thesauri; the establishment of two
clearing houses for English and non-English thesauri respective-
ly; the information given in the UNESCO bulletin; and, finally,
the efforts to cooperate with ISO in setting up regular standards
in this field.

Mr. V a r g a coined the word "superthesaurus" to mean a
system for ordering thesauri, and he refered to Mr. Rolling's
previous remarks concerning levels of compatibility.

Mr. R o l l i n g, Mr. W y s o c k i, Mr. M a l m s t e n
and Mr. S p a n g - H a n s s e n pointed to the more general
nature - es opposed to a thesaurus-like nature - of an ordering
system or a classification for thesauri.

Mr. T o m a n stressed the importance of common facets in
ordering systems for this purpose, and Mr. L e s k i suppor-
ted this view by pointing to the need of a system for common
facets and auxiliary terms to deal with e.g. geography.

As a conclusion concerning organizational problems the con-
ference recommended support for UNESCO's existing and planned
activities as regards information about the building of thesauri
and about existing thesauri.

The conference then returned to the previously postponed
Point 3; Constructional problems, in particular Point 3a; Which
elements should be included in a thesaurus?

There was agreement on the understanding of "elements", as
well as various possible presentations of a thesaurus /cf.
UNESCO Guidelines, Third Draft, as the elements, viz.descrip-
tors, non-descriptors, and relational indications, to be in-
cluded in various presentations.

Mr. J a n s e n pointed to the scheme of elements given
on p. 4 of his conference paper, and he gave priority to the

thematic group as an element in comparison with alphabetical
listing.

Mr. L e s k i, referring to p. 4 of his conference paper,
gave priority to the scheme /corresponding to the facet group-
ing in the UNESCO guidelines/ in comparison with graphic disp-
lay, and in turn priority to this in comparison with alphabet-
ical listing.

Mr. T o m a n   and   Mr. M a l m s t e n   advocated the
inclusion of all kinds of references in the alphabetical list-
ing, at least for certain practical uses, while Mr. R o l l-
i n g   pointed to certain inconveniences in burdening an
alphabetical list with, among other things, a great number of
related terms.

The indication of frequency of descriptors was mentioned as
a possible element of a thesaurus, but   Mr. S c h i f f found
this to be of little value, e.g. in indexing.

Mr. M o j ž i š e k   pointed to the usefulness of stating
the date of the introduction of a new descriptor or of elimina-
ting an absolete one to solve the problem of updating a thes-
aurus.

Mr. W y s o c k i   agreed on this point and referred to
the UNESCO guidelines sect. XIII. He found that what has been
said about presentation was in essential agreement with the
Guidelines.

Mr. T o m a n   and others would prefer another designation
than "facet grouping" for the way of presentation dealt with in
section XI /b/ of the Guidelines.

Mr. G r a v e s t e i j n   and   Mr. M a l m s t e n   ob-
ject to the concept of listing as used exclusively for alphabe-
tic lists.

Mr. M o j ž i š e k   pointed to the value of an explanatory
introduction as an element of any thesaurus, cf. sect. II of the
UNESCO Guidelines.

Mr. A i t c h i s o n   expressed doubt as to the relevance
of the remarks about computer storage, found in sect. XI /d/ of
the Guidelines.

As a conclusion concerning elements of a thesaurus, the co
ference recommended an explicit statement on the priority o

systematic displays by rephrasing the first part of sect. XI
of the UNESCO Guidelines; Third Draft in the following way:
    "XI. Presentation of Thesaurus
    It is recommended that a thesaurus be presented in one or
more systematic displays and in an alphabetical listing.
    Secondly, the conference recommends a less negative for-
mulation as regards the introduction of structure in an alpha-
betical list.
    Thirdly, the conference recommends that the wey of presen-
tation designated by "facet grouping" is designated also by
other terms in common use for this way of presentation".

- 152 -

Friday, 27th March 1970. Morning Session. Chairman: D.C.Weeks

Mr. Spang-Hanssen presented a summary of the previous sess-
ion, of which he was Chairman.

The initial topic of the session was the evaluation of the-
sauri, added to the agenda at Mr. Aitchison's request. The
Chairman specified thesaurus evaluation as having these implica-
tions:

1/ the efficiency of its function in a system - its techni-
cal qualities,

2/ the effectiveness, which is a measure of the degree to
which it is capable of fulfilling the user's needs and of describ-
ing the user's problem. When evaluation included comparison
of two or more thesauri, then it is necessary to distinguish
between thesauri that define a given domain of knowledge, and
those which are specifically system-related. In the first ins-
tance, comparison cannot be based on substantive considerations
since their content is different. In the second instance, their
content may be similar, but their system applications may dif-
fer.

Mr. A i t c h i s o n   explained his interpretation of the
matter by stating that he perceived two questions: 1/ By what
means can we compare the performance of a system employing a
thesaurus and one that does not? Is there a way of proving the
benefits in performance obtained by one of these alternatives?
Evaluation implies a comparison of different forms of thesauri.

The Chairman added that when systems with thesauri are comp-
ared with systems having none /free-text/, we must recognize
that input and processing are quite different; that the set of
rules is also dissimilar and that assessment must be made by
keeping the implications of those differences clearly in mind.

Mr. T o m a n   suggested that a useful comparison might be
made between systems employing other means, such as the Cranfield
experiment which compared various techniques. He considered the
problem to contain two distinct elements:

1/ Evaluation of other ordering systems /this would permit
evaluation of the thesaurus as a tool/ and

2/ Measurement of effectiveness /assessment of the opera-
tional effectiveness in a system/.

Mr. M a l m s t e n   added the suggestion of the anti-the-
saurus concept. Scandinavian systems received feed-back from
industrial organizations where abstracts were used.

Mr. M o j ž i š e k   stated that all systems function on
a set of rules which determine how processing is accomplished.
It is necessary to analyze the file, to evaluate the retrieval
requests, to make comparisons among these requests and thus to
evaluate the amount of noise in an operating system. He noted
the difference between an indexing language and the language
of retrieval. To reconcile their differences it is necessary to
develop a syntagmatic or meta-language. The existing difference
- one in which a system language loses its grammar - is a prin-
cipal reason for the presence of noise. Only by correlating in-
put and output languages can this problem be solved. The Chairman
stressed the gap between the efforts that have so nearly resol-
ved problems of semantics and the minor progress toward solu-
tions for the selious syntactic difficulties that impede system
effectiveness.

Mr. M a i x n e r   emphasized the importance of Mr Mojži-
šek's remarks. He added that thesauri can only be evaluated as
integral parts of a retrieval language; what when a thesaurus
states its language in a paradigmatic way and the two are then
compared, we can then observe their levels of efficiency.

Mr. M o j ž i š e k: the rules for making a thesaurus when
applied to making a system dictionary are all system-related,
producing a new language which is not that of the source /docu-
ment/, nor of the users. These languages need to develop a gram-
mar which will bring all elements into correspondence.

Mr. R o l l i n g: Apropos the Cranfield experiment, it is
essential to remember that this effort was a comparison of or-
dering systems, all placed in the same environment as opposed
to the individual, natural environments. Comparison of results
is, therefore, difficult because evaluation is usually applied
to systems as entities. Mr. Rolling urged continual internal
evaluation of a thesaurus based on failures of recall and pre-
cision. He had found that a large percentage of failures were
caused by imprecision in the thesaurus.

Mr. A i t c h i s o n remimded the group that Craufield
included two different studies. The second was able to show
significant differences in coordinato indexing.

Mr. L l o y d stated that any test or measurement was
not merely a test of the thesaurus but was in fact a system
test. He proposed that economics was a critical factor, and
in a system without a thesaurus it is possible to achieve high
recall and relevance but only at a very high cost.

Summing up the topic, Mr. L l o y d offered the view
that systems employing thesauri should be constantly evaluated
for effectiveness and economy and the thesaurus updated accor-
dingly.

·   D e f i n i t i o n o f a t h e s a u r u s:

It was suggested by Mr. T o m a n and agreed upon that
the assembly offer a description rather than a definition /to
avoid the strictures of a formal definition/ Mr. M o l n á r
offered a formal description of a thesaurus developed from a
synthesis of conference agreements. These included matters of
/1/ Order /systematic and alphabetic structure/ /2/ terms
/descriptors and non-descriptors, regretting the decision to
abandon the designation ASCRIPTOR/; and /3/ Interrelations
/preferential, hierarchical and affinitive/.

A general discussion yielded agreement that in reference
to structure, systematic order was to take precedence over
alphabetic: Mr Toman reverted to his statement at the first
meeting, when he observed that most ordering systems could be
described by these same qualities.

Four members of the conference offered written descrip-
tions which are included here:

1/ Mr. Molnár

2/ Mr. Rolling   /These are attached/

3/ Mr. Toman

4/ Mr. Poletyło

The Chairman synthesized the proposed descriptions so as
to meet these requirements:

1/ That we state the nature of a thesaurus - to what class
 . of resources it belongs.

2/ The contents are named.

3/ Its qualities are stated
   /it covers a domain/
   /it is controlled, dynamic/
   /it is arranged in systematic order/
4/ Its principal use:
   /coordinate indexing systems/
The combined description was accepted by the conference.

                              x

                           x   x

Mr. P o l e t y ł o
When we say that a thesaurus is an indexing tool, we must
add what we think about coordinate indexing. Therefore the de-
finition of a thesaurus may be as follows:
   A thesaurus is the controlled vocabulary of a coordinate-
indexing-language.
   Consequently:
   A descriptor is the preferred term in the controlled voca-
bulary of a coordinate-indexing-language.
   A non-descriptor is a forbidden term in the controlled vo-
cabulary of a coordinate-indexing-language.
   Without adding "coordinate", the definition will be too
broad. The term "coordinate" is a feature that distinguishes a
descriptor language from all indexing languages, in contradis-
tinction to precoordinate-indexing-languages as classification
and subject-heading languages.
   The term "controlled" distinguishes a descriptor language
from coordinate-indexing-languages and from a uniterm language
/which is uncontrolled/.

   Mr. I. M o l n á r
   After a 4-day conference, after many friendly and halpful
discussions, after many agreements, approximate agreements and
disagreements too, we have not reached an agread definition of
our fundamental topic, although discussions have taken more
than 20 hours.
   Could such a meeting of experts, practical creators of in-
formation systems and thesauri, not find a name for their com-
mon beautiful son?

In contrast to this, every participant in the conference
has his own definitions for thesauri. It is a pity that these
definitions reflect as many points of view as the number of
participants.

Well, in this difficult situation, in this uncomfortable
richness of definitions we have the duty to find the common
characteristics of them all. This could be the first approxima-
tion of the problem. Forgive me if I begin with a rather gene-
ral definition of my own. This is as follows:

1. **A thesaurus is in general an ordered collection of terms
which also includes their inter-relation**

This includes no information on either the principle of or-
dering, or the nature of terms, or the branches of science, or
the types and nature of inter-relations between the terms inclu-
ded, nor the task of the thesaurus. This definition is of such
a general validity that everyone can agree with it. Everything
it includes is inevitable, a "sine qua non" for a thesaurus. But
this determination does not include some elements which must be
further investigated.

The definition in this form is able to cover all thesauri
from Roget's to the Euratom thesaurus.

I hope we can be agreed on the basis of minimum programme.
If so, we are able to move on in the direction of deeper deta-
ils, leading to a more detailed definition.

a/ The _first_ element in my definition which requires further
   analysis is the expression _ordered_. This word includes the
   problems of the structure of a thesaurus.
   We have already found agreement in relation to this expres-
   sion. The material of a thesaurus can be ordered alphabeti-
   cally or systematically, but it is advisable to prepare both
   an alphabetical and a systematical ordering.
   By systematical ordering we understand the hierarchical or
   graphic display of terms.

b/ The second element which must be investigated is the expres-
   sion _terms_.
   We already have a common opinion on this question too. Terms
   are defined as consisting of descriptors and non-descriptors.
   I really think it regrettable that the term _ascriptor_ is

dead in "statu nascendi"/. But we are in agreement, I see,
on the problem of auxiliary terms, too. These seem to be of
a descriptor character.
/The definition of descriptor is not included in the defi-
nition of a thesaurus. This is for the sake of brevity/.
c/ The third element to be detailed is the expression inter-
relations. We tend to egree on the recommendations of the
UNESCO Guidelines and so we have three groups for inter-re-
lationships of terms.
These are

                  1. preferential
                  2. hierrrohical and
                  3. affinitive

relations. The common name of 1. and 3. is "non-hierarchi-
cal inter-relationships".
As a consequence of these agreements-in-detail, the second
approximation of the definition could be as follows:

   2. A thesaurus is in general an alphabetically or/and sys-
tematically ordered collection of descriptors and non-descrip-
tors having hierarchical and/or non-hierarchical inter-rela-
tionships.

   This second definition is somewhat more precise and of a
more specialized character.
   There is an unnecessary element in the definition.The third
approximation is directed to words the elimination of this ele-
ment, finding a better corresponding word instead. This unneces-
sary element is the expression in general.
   This elimination can be realized only by determining the
precise purpose of the thesaurus.
   In the last three days we have talked a lot about the va-
rious possibilities and directions of the further development
of thesauri. We have verified the science-developmental func-
tion, its science-organizing power, we mentioned the linguistic
possibilities, eto. But we gave the greatest part of our dis-
cussions to the use of thesauri in indexing work, information
storage and retrieval, as well as to the standardization of in-
formation queries and to the further development of soientifio
information.

It is well known that there is a general and essential correlation between function and structure in the living world as well as in dead material. And this correlation is valid in the case of thesauri, too.

A thesaurus, the task of which is to introduce someone into the world of science, must have a given structure which makes this possible. This structure can be a hierarchical one. Another thesaurus intended as an effective tool for indexing documents must be of alphabetically arranged structure, because this is the only way an indexer can do his job.

On the basis of this explanation it is possible to realize the third approximation for a definition of thesauri, but only when postulating its specific task.

We are information specialists. Thesauri are very important tools for our fundamental activity. Therefore we must investigate only <u>one side</u> of the whole problem of thesauri,and this is the side of information activity.

After this statement I feel it possible to define a thesaurus in a more exact form:

3. <u>A thesaurus as a tool of information systems is an alphabetically and/or systematically ordered collection of descriptors and non-descriptors having hierarchical and/or non-hierarchical inter-relationships.</u>

This definition seems to correspond to our information tasks, though it includes no statements in connection with traditional or computerized uses. But it would be incorrect because of the relatively great number of existing traditional systems.

Other thesauri for other purposes can be otherwise defined.

Mr. R o l l i n g

A thesaurus is a controlled but dynamic vocabulary of semantically related terms offering comprehensive coverage of a specific domain of knowledge.

Its main use is in the subject characterization of documents and queries in information storage and retrieval systems based on concept coordination.

Its principal elements are descriptors, non-descriptors, terms and relationship indicators.

It generally comprises one or more systematic displays and one or more alphabetical listings.

Mr. T o m a n

A thesaurus is a system of controlled terms used for characterizing the content of documens /information/ in storage and/or retrieval systems. It is characterized in its dynamics by a weak hierarchy, by preferring postcoordination to precoordination. In contrast to classification systems, it displays affinitive terms but does not use notation and categories.

Friday, 27th March. Final Session-11.45 a.m. Chairman: K.Leski

The session was devoted to the problem of formulating con-
clusions embracing the results of the whole conference.   The
following conclusions were taken:

I. A thesaurus is a lexical tool of information retrieval
systems. It consists of a controlled but dynamic vocabulary of
semantically related terms. This vocabulary,which comprehen-
sively covers a specific domain of knowledge, is a systemati-
cally and alphabetically ordered collection of descriptors,non-
descriptors /auxiliary terms/ as well as indicators of their
relationships both hierarchical and non-hierarchical.

Unlike classification systems a thesaurus does not neces-
sarily use notations and categories.

Its main use is in the subject characterization of documents
/information/ and queries in systems based on concept coordina-
tion.

When discussing the question of the display of descriptors,
the importance of the systematic display, whether in thematic,
hierarchical or graphical form, or all of them together,   was
stressed. It was recommended that the UNESCO Guidelines put the
systematic display before the alphabetic in their wording.

II. The main role of a thesaurus is to ensure that indexing
and retrieval of documents can be affected with a maximum of
precision and completeness.

In addition, the existence of a thesaurus covering a subject
field tends to stabilize the terminology of this field.

In particular:
  - Thesaurus work leads to the formulation of more precise
    definitions of inconsistently-used terms.
  - New terminology is properly defined and given its prefer-
    red place in the semantic structure of the vocabulary.
  - Edit s could use thesauri in advising authors to use
    cor ct terminology in their publications /particularly
    in titles, which are used for KWIC and other indexes/.

- Editors could use thesauri for preliminary logging of
  articles or abstracts, thus facilitating the work of
  subsequent scanners and indexers.
- Dictionaries should include mention of preferred termi-
  nology in the major thesauri.
- Thesauri constitute the starting point for elaboration
  of the more complex lists of words used in free-text
  processing.

Systems employing thesauri should be continually scrutiniz-
ed and evaluated in terms of recall, relevance, and econo-
mics: and the thesauri should be updated accordingly.

III. Three types of factors influence the elaboration of
thesauri:
   1. Factors related to subject area and language
      - volume of literature to be covered
      - language distribution of this literature
      - redundancy /or lack of precision/ of terminology
      - overlapping with fields already covered
   2. Factors related to the user population
      - degree of precision required
      - degree of completeness required
      - response volume required
   3. Factors related to system methodology
      - equipment used /degree of mechanization/
      - storage media used
      - file organization
      - search logic

IV. The general opinion was that it was impossible to build
a universal thesaurus in the sense of the detailed UDC univer-
sal system, but that if the need is felt for a superstructure
covering the existing special thesauri it should be a very ge-
neric classification scheme in the sense of the recommendation
made during the UNISIST control committee meeting in December
1969. It was suggested to the UNESCO delegate to provide,when
constructing a general scheme, for lists of auxiliary terms and
common facets from which the special thesauri could eventually
draw the terms needed for the description of the form of docu-
ments and such terms which are common to many thesauri.

In order to obtain a worldwide information system, three
levels of ordering systems are necessary:
1/ Mission- or discipline-oriented thesauri serving on a
   specific level for indexing and retrieval purposes
2/ General thesauri assuring compatibility between speci-
   fic thesauri in related fields
3/ An ordering system displaying the different branches
   of human knowledge at a very general level.

V. The problems of multilingual thesauri were discussed
exhaustively. It was agreed that there are two approaches to
the building of multilingual thesauri: simultaneous construc-
tion of thesauri in different languages, and translation of a
thesaurus in one language into thesauri in other languages.
Similarly, there are two methods of use: simultaneous use of
the thesaurus by two or more language groups, and transfer of
a data base from one language to another. It was recommended
that UNESCO should take account of the deliberations of the
meeting in preparing their guidelines for multilingual thesa-
uri.
    The conference agreed in principle with the UNESCO Guide-
lines, Section VI - "Selection of descriptors" and Section IX -
"Methods of entering descriptors in the Thesaurus" and agreed
to provide UNESCO with their own comments.

GUIDELINES FOR THE ESTABLISHMENT AND DEVELOPMENT
OF MONOLINGUAL SCIENTIFIC AND TECHNICAL THESAURI
FOR INFORMATION RETRIEVAL

United Nations Educational,
Scientific and Cultural Organization[x]

E x p l a n a t o r y   s t a t e m e n t

At a time when the establishment of a World Science Infor-
mation System is being seriously proposed[xx] it is advisable to
rememember that the viability of any world system depends first
and foremost on compatibility between its component parts.

These guidelines for the establishment and development of
monolingual scientific and technical thesauri for information
retrieval are published in an attempt to lay the basis for
compatibility, both at present and in the future, of thesauri
that are being elaborated simultaneously in most of the discip-
lines of science, basic as well as applied.

They are, therefore, directed to all those who in the cour-
se of their career come into contact with thesauri, either as
users or as thesaurus compilers.

The first draft of the guidelines was prepared by the Unes-
co Secretariat. The third draft and this version were subsequ-
ently reviewed and studied by eminent and competent individuals
and organizations and the relevant additions or corrections

---

[x] SC/MD/20 - Paris, 6 July 1970 - Original: English
   These guidelines were specifically drafted for the English
   language and when applied to monolingual thesauri in other
   languages they should be modified to take into considera-
   tion the attributes and uses of that particular language.
[xx] Joint Unesco-ICSU study on the feasibility of a World Scien-
   ce Information System /UNISIST/ Final Report. Unesco, Paris,
   1970.

were made. Thus, these guidelines were presented to, and dis-
cussed by, the International Conference on the General Princip-
les of Thesaurus Building in Warsaw, March 1970. The proposed
changes are included in this version /1/. May our collaborators
accept anonymity along with our gratitude.

Fourteen guidelines are presented: the first four are of a
general nature, the following seven deal with the establishment
of thesauri, and the final three relate to the development of
thesauri. Examples, where appropriate, are given on the right-
hand margin of the text.By the word "thesaurus", as used in the
present text, is meant a controlled and dynamic vocabulary of
semantically and generically related terms which comprehensively
covers a specific domain of knowledge. This vocabulary is a sy-
stematical and/or alphabetical collection of descriptors, non-
descriptors /auxiliary terms/ as well as indicators of their re-
lationships. Unlike classification schemes, the vocabulary does
not necessarily use notations and categories.

A descriptor is an authorized and formulized term or symbol
in a thesaurus, used to represent unambiguously the concepts of
documents and queries /2/.

Thesauri should be based on concepts and relationships which
are internationally acceptable. Original and translated thesauri
already exist in most of the major vehicular languages used in
science and technology today. It is rare that any particular
word can be translated univocally into another language without
losing some shade of meaning in the process, but it is hoped that
the application of these guidelines to monolingual thesauri will
diminish the enormous difficulties encounterd in the establish-
ment of thesauri in different languages. These guidelines were
originally drafted in English and when applied to monolingual
thesauri in other languages, they should be modified to take in-
to consideration the attributes and uses of that particular lan-
guage /e.g. number of descriptors, VIII b/.

Thesauri can be used in many ways, and the structure of a
thesaurus is intimately related to its proposed utilization. A
thesaurus can be used merely as a word association list for hel-
ping indexers, or it can be conside ed as a transformation of the
natural language into the informati n language /3/.

Modern techniquea in information acience are nearly all
based on the use of electronic computers and it is in this
connexion that the use of thesauri is rapidly proliferatiug.It
is this rapid proliferation which has brought the need for in-
ternational guidelines to light and it waa for this reason,too,
that Uneaco recently encouraged /helping in the establishment
of one/ the work of two clearing-houses dealing with thesauri.
Theae clearing-housea are located at the Bibliographic Systems
Center, School of Library Science, Case Western Reserve Univer-
aity, Cleveland, Ohio 44 106, United Statea of America, and at
the Centralny Instytut Informacji Naukowo-Technicznej i Ekono-
micznej, Al. Niepodległości 188, Warsaw, Poland for English and
languagea other than English respectively.

General

I. Adviaability of a pilot run

Refore eatablishing a thesaurus on a definitive basis it is
strongly recommended that a practical test, baaed on a reatric-
ted number of documents dealing with a small area of the domains
to be ultimately covered, be carried out. This pilot run, based
on tentatively atructured terms, should ahow up the more adequ-
ate methods of descriptor aelection and thesaurus display   ap-
plicable to the case under consideration. The reaulta of this
teat should be critically commented upon by as many people as
feaaible, including information acientista and indexera as well
as aubject specialista and users.

II. Neceasity of a deacriptive in-
troduction to the theaaurua

No theeaurus should be preseuted without a comprehensive in-
troduction which states clearly the purpose and structure of the
thesaurus, and the domains covered by it. The rules followed in
its establishment ahould be p aaented in a condensed form.This
is particularly true of the me ;hods aud aources used in   the
selection, form and avoidance of ambiguity of the descriptors
/aee VI, VII, VIII/.The method of presenting the theaaurus as well as

the rules for alphabetization and punctuation, whenever appli-
cable, should be explicitly stated.

Most important of all, the rules for using the thesaurus
and its limits of applicability should be elucidated and illu-
strated by means of examples, where appropriate.

Users should be invited to contribute comments and sugges-
tions for the improvement of the thesaurus, and to inscribe
themselves on the mailing list for future editions of or addi-
tions to the theaaurus. The proposed system for developing and
up-dating the thesaurus should be explained; the date of the
present, and estimated appearance of future, editions or addi-
tions to the thesaurus should be given.

The total number of deacriptors, non-descriptors /4/,iden-
tifiers, /see VI/, hierarchical chains /see X/b// and related
concepts /see X /c// should be itemized.

III. N e c e s a i t y   o f   i n d e x e s

Every theaaurus, regardless of its mode of presentation
/see XI/ shouli contain an alphabetical union list of each in-
dividual unstructured term /5/ whether isaued aeparately as a
supplement or together with the main theaaurus as an annex.Per-
mutation indexea may also be used.

It may be useful in the case of multidisciplinary thesauri
to present, in addition, indexea in which the desoriptors are
grouped by discipline.

IV. N o t i f i c a t i o n   o f   i n t e n t

The appropriate clearing-house /see above/ should be noti-
fied of the intention to construct a thesaurus, aa well as when
the thcsaurus is first published or disseminated. This informa-
tion ahould be channelled through the national organization
dealing with thesauri, where aud when such an entity exists.

The same applies for further editions. If at all posaible,
a copy of the thesaurus, complete with the introduction and
indexea should be sent to the clearing-house in question. The
fact of notification should be mentioned in the introduction.

Establishment

V. Check with clearing-
   house to avoid dupli-
   cation

   Before commencing work on the establishment
of the thesaurus, it is advisable to ascertain
whether others covering that particular domain
or a neighbouring one are available.

   This is best done by addressing a query to
the two clearing-houses mentioned above. It may
be found advisable to go ahead with the compila-
tion of a particular thesaurus in spite of the
existence of a similar one. In this case the re-
asons for proceeding and the differences with
the earlier thesaurus should be clearly stated
in the introduction.

VI. Selection of descrip-
    tors

   The selection of descriptors should begin
only after the general structure of the thesa-
urus has been agreed upon. It should be carried
out, preferably, by people who have both a good
knowledge of the subject to be treated, and
previous experience in indexing or classifica-
tion. The use of internationally recruited te-
ams for the construction of thesauri is to be
encouraged since it widens the cumulative lin-
guistic experience which goes into the building
of the thesaurus. The methods of selecting des-
criptors vary according to the proposed structu-
re of the thesaurus /alphabetical listing,sys-
tematical listing /6/, graphic display /6/,
see XI/, the purpose for which the thesaurus
will be used /e.g. for manual or mechanical
retrieval, only for indexing, or as a secondary
tool/ and the background to the project /gra-

dual build-up to mechanical processing, intro-
duction of a new domain e.g. interdisciplinary
areas for which no previous classification   Acoustical
schemes existed, existence of well-defined group   Holography
of users and subject specialists, extensive   Brain
literature/.   Research

Descriptors, in general, consist of terms
related to discrete concepts encountered in
the subject field under consideration and in
pertinent marginal areas. A more specific class
of thesaurus terms /7/ known as "identifiers"
may sometimes be used.

Descriptors should succinctly summarize
concepts in as few words as possible,preferably
one. Grammatical connexions such as preposi-
tions or articles should be avoided whenever
possible.

Identifiers constitute a special type of
thesaurus terms /8/ which are not reciprocally
cross-referenced /see XI/ and which serve the
purpose of providing additional indexing depth.
For instance, identifiers might include indi-   IRELAND
vidual trade names, geographical locations,   NT DUBLIN
equipment, nomenclature, code names etc.   DUBLIN

Since they are not reciprocally cross-re-
ferenced, identifiers need not necessarily
appear in the thesaurus display, but may be
listed separately, in addition to appearing
in the Union List /see III above/.

Four distinct steps intervene in the se-
lection of descriptors: collection, verifica-
tion, evaluation, and choice.

/a/ Collection
It is _most impossible to make a compre-
hensive c llection of candidate descriptors
by tinking of an alphabetical list. By en-
visaging descriptors in groups, thought asso-
ciations between them give rise to many candi-

dates. Potential users end subject specialists
as well as internationally or nationally stan-
dardized technical dictionaries should be
consulted; terms should be chosen from the cu-
rrent literature; existing word lists or class-
ification schemes should be culled and may be
expanded or compressed appropriately.Scientific
and technical dictionaries and glossaries, both
multilingual and monolingual constitute a pro-
lific source of descriptors /see page 182/.

/b/ Verification

With all methods of assembly,the authenti-
city of the selected descriptors should be
verified by consulting dictionaries, other in-
dexing or standardized vocabularies, current
usage in the literature and aspecially the opi-
nion of subject specialist. Obsolete termin-
ology should not be included, or if so only as
forbidden terms /see X /a/ below/.

One of the more appealing attributes of a
thesaurus is its ability to assimilate immed-
iately the neologisms and special jargon that
proliferate in expanding fields of.basic   and
applied research. Full advantage should be taken .
of this facility in combination with the use
of scope notes /see VII /c/ below/ and   cross-
references.Special care should be taken with
terms whose connotations have changed with the
passage of time, or whose meaning changes from
country to country. If overlapping terms have
to be included the appropriate cross-reference   BILLION/10 EXP 9/
/see X below/ should be employed.   BILLION/10 EXP 12/

/c/ Evaluation

In evaluating the utility of candidate
descriptors, reference should be made to their:
1. frequency as encountered in the literature
or in the existing stocks of information/9/;

2. anticipated incidence in retrieval inquiries;

3. relationship to descriptors already accepted;

4. appropriateness and authenticity as current termino-
logy in the discipline concerned; 5. effectiveness and
expediency in connoting and denoting the particular con-
cept. None of these factors should be considered inde-
pendently and particular attention should be paid to areas
of peripheral interest where the exhaustivity and speci-
ficity required of the descriptors are not the same as
for the core subject.

/d/ Choice

In all cases, descriptors should be selected for in-
clusion in the thesaurus on the basis of their estimated
effectiveness for retrieval purposes and their measureable
significance in the material to be indexed.

VII. M e t h o d s   o f   a v o i d i n g
    a m b i g u i t y

In compiling a thesaurus, difficulties are encounter-
ed with descriptors which have more than one accepted
meaning or whose meaning in a given context is different to
that commonly encountered. In such cases the required
meaning may be brought out by the use of the following me-
thods:

/a/ Compound expressions

Although descriptors are preferably self-contained,
single term concepts, the use of modifying expressions to
make clear the different meanings associated with a given
term is necessary in certain cases. For the method of en-
tering the resulting compound expression,/see IX/a/ below/.

LATENT
HEAT

/b/ Qualifiers for homonyms

The various forms of homonyms may be distinguished by
the use of qualifying expressions placed between paren-
theses immediately after the homonym. Other homonyms should
not be used as parenthetic qualifiers.

BEAMS
/ELEC-
TROMAG-
NETIC/
BEAMS
/STRUC-
TURAL/

/c/ Scope notes

A scope note is a brief explanation which may
accompany the descriptor in the thesaurus display,

but does not form part of the descriptor.It indi-
cates the way in which the descriptor should be
used; it need not necessarily consist of a dic-
tionary definition. Scope notes are sometimes
used to restrict the usage of e descriptor. They
should always be used in connexion with abbrevia-
tions and acronyms /see VII /d/ below/.

It is recommended that either /a/ and /c/ or
/b/ and /c/ above be.used together in a single
thesaurus. Methods /a/ and /b/ should be mutual-
ly exclusive and never be used in one and the
same thesaurus.

*DOCUMENTA-
TION
The process
of storing
and retriev-
ing informa-
tion in all
fields of
learning.
*DOCUMENTA-
TION
The volume
of documents
assembled or
available.
*DOCUMENTATION
The title of s
family of pub-
lications.

VIII. F o r m   o f   d e s c r i p t o r s

/a/ Word form

Once it has been decided to include a given
term in the thesaurus, care should be taken to
ensure that the word form used adequately con-
veys the exact meaning intended.

/i/ Spelling: the most widely accepted
spelling of the word should be used. Cases
arise, particularly in English, due to varying
usage on different sides of the Atlantic,where
more than one spelling of a word is accepted,in
which case both forms of the word should.be inclu-
ded in the thesaurus. In these cases the prefe-
rential cross-reference should be.employed /see
X /a/ below/. Alternatively, a well-established
dictionary can be chosen to act as arbitrator
whenever this problem arises.

SULFUR
SULPHUR

/ii/ Translation: many current technical
terms have arisen by translation from other lan-
guages, but sometimes a modern foreign language

X For instance, in three different thesauri. If these three meen-
   ings were in the same thesaurus, they would require qualifiers
   in order to make them unique.

or Latin term is incorporated into the spe-
cialized vocabulary for a particular subject.
When both the foreign language term and its
putative translation coexist, they should   BRAKING RADIATION
both be included in the thesaurus and cross-   BREMSSTRAHLUNG
referenced preferentially.

   /iii/ Transliteration: the problem is
further complicated when the foreign langu-
age in question is written in a different
alphabet. This is particularly true in the
case of identifiers /see VI above/. The
transliteration standards[x] recommended by
the International Organization for Standardi-
zation should be used whenever applicable.
Wherever a choice exists the transliteration   SATELLITE
which does not employ diacritical marks   SPUTNIK
should be selected /see /e/ below/.

/b/ Noun form

   The descriptor should be in the form of   The gerund in
a noun or that part of the verb which is   English
grammatically equivalent.

/c/ Number

   In general, the plural form should be   FORCES
used for descriptors, particularly when ge-   HEATING
neric terms are involved. The singular form   PALYNOLOGY
is used for specific material or property   TEAK
terms, process terms, proper names and dis-
ciplinary areas. Sometimes the singular and   WOOD
plural forms of a word denote different con-   WOODS
cepts; in this case both should be entered.

/d/ Abbreviations and acronyms

   Abbreviated word forms should be used
only when their meaning is international‍ly

---

[x] See page

established. Both abbreviated and unabbreviated
forms should be displayed and cross-referenced
preferentially. Sometimes the necessity   for   UNESCO
limiting the length of the descriptor /see /e/   United Nations
                                                    Educational,
below/ entails the use of less well established   Scientific and
abbreviations. In all these cases a scope note   Cultural
                                                    Organization
/see VII /c/ above/ should be appended.

The above remarks also apply to acronyms.

/e/ Character set

Since the majority of scientific and tech-
nical thesauri now being established will prob-
ably be used in connexion with electronic com-
puters, it is advisable to use only the upper
cese format for the descriptors. Diacritical
marks should be avoided for the same reason.

The need for these restrictions will prob-
ably disappear in the near future, as the fruits
of technical advances become more widely dis-
tributed, and computer manufacturers pay more
heed to the exhortations of information scien-
tists to lower the costs of peripheral equip-
ment.

As mentioned in /d/ above, the eventual use
of a computer may entail the limiting of the
number of characters that a descriptor may have.

/f/ Special characters and numerals

The only special characters allowed in des-
criptors are left and right parentheses and
unavoidable hyphens. /Fullstops may sometimes
be used /see IX /b/ below/. Any other non-
alphanumeric symbols should be confined to scope   MERCURY
notes, always within the limits of machine cha-   /PLANET/
racter availability. If the descriptors contain
numeric elements, arabic numerals should be used.
The position of the numerals should follow nor-
mal usage.Rules must be established for the
treatment of subscript and superscript numerals.

In the particular case of data retrieval
thesauri, the stroke /"/"/ may sometimes be   EH/M
found necessary.

IX. M e t h : d s   o f   e n t e r i n g
    d e s c r i p t o r s   i n   t h e
    t h e s a u r u s

/a/ Syntax

Compound expressions consisting of two
or more words should be listed preferably by
direct entry i.e. not artificially inverted.
This is especially true for /10/descriptors:
for forbidden terms, this recommendation may   ELECTRICAL
be relaxed. Evidently this does not apply   POWER
                                                  not
when a permuted or key word in context type   POWER,
of multiple entry is used. Inverted entries   ELECTRICAL
may be used provided they are preferentially
cross-referenced /see X /a//.When a qualifier
between brackets /see VII /b/ above/ forms
part of the descriptor it is advisable to enter
the qualifier on its own with a preferential
cross-reference to the complete descriptor.

/b/ Punctuation

Punctuation marks should not appear in
the descriptors. As stated in VIII /f/ above,
the only non-alphanumeric symbols normally   LIGHT-
allowed in the descriptors are left and right   SENSITIVE
parentheses. Fullstops should only be allowed   DEVICES HIGH-
when, due to a limit on the length of the   VOLTAGE
descriptor, a word has to be truncated. Hy-   PYLONS
phens should only be used when their omission
would alter the intended meaning of the de-
scriptor. Commas, colons and apostrophes should
be excluded since they are not necessary to
convey the meaning of the terms. Where punctua-
tion marks are omitted, it is advisable to in-
clude them in full in the scope notes.

/c/ /i/ Specialized vocabularies

Certain fields have highly specific systems of nomenclature, or well-established standardized technical vocabularies. Whenever an internationally agreed nomenclature exists, it should be used.

/ii/ Specific names

The proliferation of unrelated specific names would tend to convert the thesaurus into a simple list of identifiers which would be aelfdefeating. It is therefore recommended that the names of unrelated specific entities be avoided aa much as possible.

/iii/Specific items

Descriptors representing generic, functional or structural concepts can be co-ordinated to denote specific items, while by retaining the property of being cross-referenced, they fulfil the structural needs of thesaurus elements.

/d/ Alphabetization

Where appropriate, one of the following alphabetization methods may be followed:

/i/ letter by letter

/ii/ word by word

/iii/ computer sort

The selection of the method of alphabetisation depends on all the factors affecting the thesaurus under construction. i.e. the size and structure of the domains covered by the thesaurus, the availability of machine processing, the kind of hardware available, etc. In all cases the alphabetisation rules should be clearly and explicitly drawn up before any kind of ordering is attempted.

/e/ <u>Synonyms and quasi-synonyms</u>

It is rare that two or more candidate
descriptors can be considered as true
synonyms. When one candidate descriptor·
must be searched every time that another
is searched, they may be treated as syno-
nyms. Descriptors that overlap significan-
tly or represent different aspects of the
same property may be considered quasi-
synonyms. Antonyms should be similarly
treated. When all the synonyms, quasi-
synonyms or antonyms are included in the
thesaurus display the preferential cross-
reference should be used /see X /a/below/.

COLUMBIUM/NIOBIUM

HEREDITY/GENETICS

HARDNESS/SOFTNESS

X. I n t e r r e l a t i o n s h i p s
    b e t w e e n   d e s c r i p t o r s

The most important function of a the-
saurus is to serve as a tool for informa-
tion retrieval. Therefore it should bring
into evidence the interrelationship
between individual descriptors. These can
be expressed by several means. If codes
are used to indicate these relationships,
their meaning should always be made clear.

These interrelationships are of three
types: preferential; hierarchical; affini-
tive. All three have the property of rec-
iprocity, i.e. when two or more descrip-
tors are related in any way, reciprocal
entries are required. /Identifiers /VI
above/ are the only exceptions to this very
important rule./

This is nece·ssary of the homogeneity
of the the· .rus and for "book-keeping"
purposes.

/a/ Preferential

This reference is employed to refer from a forbidden term to /11/ a descriptor and vice versa. It is used when the meaning of descriptors overlaps substantially; where different spellings of the same word exist; for synonyms, quasisynonyms and antonyms and, in general, wherever a choice has been made between a number of descriptors, all of which are included in the thesaurus display.

Common codes in English are: use/ /includes use/USE// /used for /UF/

ALCOHOLS
  USE AL-
  KANOLS
ALKANOLS
  UF AL-
  COHOLS

/b/ Hierarchical

Hierarchical relationships are used to exhibit relative degrees of specificity within a category of descriptors all of which belong to a particular generic group. This relationship is not based upon the possible use or application of an entity, but on the position of the descriptor within a given class of concepts. Note that certain terms may be members of more than one hierarchical chain. Where any hierarchy has more than two levels the cross-references for all levels should be completed for each descriptor. The kinds of hierarchical relationships which it is desirable to indicate depend on the structure of the subject field of the thesaurus. In general, all concepts which are sub-divisions of a broader concept should form part of a hierarchical chain.

Common codes in English are: broader term /BT//narrower term /NT/ specific to/generic to

CALCULUS NT INTEGRAL CALCULUS

INTEGRAL CALCULUS BT CALCULUS

Genus-species in zoology

Whole-part

Subordinate concepts

/c/ Affinitive

The affinitive relationship is employed to refer from a descriptor to others that are closely related in concept but are neither consistently hierarchically nor preferentially related. This relationship may be based on usage, application, physical proximity, etc.

Common codes in English are: related term /RT/ also see EDUCATION RT LEARNING LEARNING RT EDUCATION

XI. Presentation of thesaurus

It is recommended that a thesaurus be presented in one or more systematical displays and alphabetical listings /12/.

/a/ S y s t e m a t i c a l   l i s t i n g

Systematical listing /13/ refers to that form of thesaurus
display in which descriptors are first of all grouped in general
class categories within each of which the interrelationships
between the descriptors, particularly the hierarchical rela-
tionships, are as self-contained as possible. Full use should
be made of recorded experience in the field of classification
when establishing the membership of the various facets.

Some descriptors may appear in more than one category but
this should occur only when either the descriptor is accompa-
nied by a parenthetical qualifier or when cross-references are
used.

Thesauri that are presented in this way should always con-
tain an alphabetical listing of all the terms included in the
thesaurus /see III above/.

Systematical listing /14/ is probably better for very spe-
cialized scientific and technical fields than for interdis-
ciplinary areas.

Combination of this type of display with /c/ below gives
rise to a kind of structured alphabetical list which probably
combines to the fullest extent the advantages of both.

/b/ Graphic display

Perhaps the most subtle mode of presentation of thesauri
is to display the descriptors and the relationships between
them graphically. Although this can be done multi-dimensionally,
for instance by taking two dimensions for each facet of a
multi-faceted thesaurus, the more current methods are two-di-
mensional.

· One such system consists of arranging the descriptors in
semantic groups, assigning a gridded sheet to each group and
giving fixed positions to each descriptor with respect to the
horizontal and vertical axes, thus defining co-ordinates.

Interrelationships between descriptors are then shown by
means of arrows. Associative relationships are denoted by bi-
directional arrows. Hierarchical relationships are shown by
unidirectional arrows always pointed to the more specific des-
criptor. Preferential relationships may be indicated by brac-
kets with the arrows leaving or arriving at the preferred term.

It is understood that a descriptor may belong to several
groups. The optimal size of each group appears to lie between
30 and 40. As before, an alphabetical listing should be given
in annex showing the semantic group/s/ to which each descrip-
tor belongs. Which mode of presentation is selected will de-
pend on the use to which the particular thesaurus will be put.

The latter two types of display land themselves more easily
to translation. A rather particular type of thesaurus is the
following.

/c/ Alphabetical listing

The great advantage of an alphabetical listing is that
the introduction and correct positioning of new descriptors is
very easy. On the other hand, it is extremely difficult to
introduce structure into a strictly alphabetical list. For
instance, synonyms come more readily to mind if we think of a
particular category as a whole rather than individual descrip-
tors plucked at random from an alphabetical list. It should be
remembered that a particular alphabetical order is only appli-
cable in one language. Parmuted alphabetical lists may also be
used /15/.

Development

XII. Periodic   verification   of   us~-
     fulness   of   individual   descrip-
     tors

At least for the first few years, if not permanently, after
the establishment of a thesaurus, a check should be kept on the
frequency with which particular descriptors are utilized, both
for indexing and retrieval purposes. Periodic verification should
ensure that certain descriptors neither interfere with, nor
duplicate one another. On all occasions in which a search does
not locate the desired information or the amount of information
suspected of being in the collection, a critical appraisal of
the descriptors which were, or should have been used, ought to
be carried out.

XIII. Elimination of descriptors

If it is found that any descriptor is being used very infrequently, care should be taken to ensure that the infrequency of usage is not due purely to the lack of documents related to that particular concept. It may either be eliminated from the thesaurus or replaced by another more common term. Complete elimination should occur ideally only when that particular descriptor has never been used, either for indexing or retrieval purposes. The use of a preferential relationship to indicate where the replacement has been effected is more practical.

The inverse is also true: if too many indexed materials are assigned to the same descriptor, its specificity is lost, its application has become too general and the breaking-down of the concept should be considered.

If a preferential relationship is not used, the date of introduction of a new descriptor into the thesaurus should be noted since, prior to that date, indexers were not authorized to use that term.

The procedure to be followed when a particular descriptor is over or under used depends to a certain extent on the search strategy employed in retrieval. If the least specific descriptor is searched for last, it may not be worthwhile to eliminate it.

XIV. Choice of new descriptors

Indexers and users should constantly be on the look-out for new candidate descriptors which may represent either new concepts or different facets of old concepts. If possible, the descriptor should be used on a trial basis by indexers for some time before becoming a definite addition to the thesaurus.

The frequency of occurrence of such candidate descriptors both as indexing and retrieval terms is a good indication of their future usefulness. If it is decided to add a new descriptor, the interrelationships with all the pre-existing descriptors should be identified and introduced in the appropriate places.

Definite additions should not be introduced singly as this
csuses confusion among the users of the thesaurus. New descrip-
tors should be saved up and introduced by batches, either as
"additions to the thesaurus" or on the occasion of a new edi-
tion of the thesaurus. This does not preclude their use by in-
dexers. There should exist a central authority which examines
all the suggestions received and issues a final verdict on the
acceptability or otherwise of the possible new additions.

It should always be remembered that a thesaurus is never
completed, its size and shape being a function of time.

N o t e: Numbers in brackets from /1/ to /15/ signify pla-
ces in which the final text "Guidelines for the Establishment
and Development..." differs slightly from the text of Project 3.

**List of ISO recommendations
related to these Guidelines**

ISO/R 9   "International sytem for the transliteration of
             slavic Cyrillic characters" 2nd edition.
ISO/R 233   "International system for the transliteration of
             Arabic characters".
ISO/R 259   "Transliteration of Hebrew".
ISO/R 704   "Naming principles".
ISO/R 843   "International system for the transliteration of
             Greek characters into Latin characters".
ISO/R 860   "International unification of concepts and terms".
ISO/R 919   "Guide for the preparation of classified vocabu-
             laries".
ISO/R 1087   "Vocabulary of terminology".
ISO/R 1149   "Layout of multilingual classified vocabularies".
ISO/DR 1951 "Lexicographical symbols, particularly for use in
/Draft/   classified defining vocabularies".

The above documents are available either from the Headquar-
ters of ISO /International Organization for Standarization/,
1 rue de Varembe, Geneva 20, Switzerland.
or from: the corresponding National Standards Organizations
of the member countries of ISO.

Sources for dictionaries and glossaries:

Bibliography of interlingual scientific and technical dic-
tionaries, 5 ed. Paris, Unesco, 1969. 250 p.

Bibliography of monolingual scientific and technical glo-
ssaries, Vol.I: National Standards, 1955, 219 p. Vol. II: Mis-
cellaneous Sources, 1959, 146 p. Paris, Unesco.

/Supplements published in Babel, International Journal of
Translation published by the International Federation of Trans-
lators with the assistance of Unesco. Avignon, France./

Bibliographic Bulletin of the Clearinghouse at CIINTE.1969,
Warsaw, CIINTE, 1969, 140 p. /Annual supplements are planned./

Bibliographic Systems Center Subject Index. Case Western
Reserve University, Cleveland, Ohio, U.S.A., 1969. /Computer
print-out./

Some national standards institutions publish extensive
unilingual and sometimes bilingual technical vocabularies.

QUESTIONNAIRE ON PROBLEMS THOUGHT DESIRABLE TO RAISE
BEFORE THE MEETING

1. D e f i n i t i o n   o f   a   t h e s a u r u s

   a/ what does "thesaurus" mean?
   b/ which structural elements /semantic, syntactic etc./
      should be included in order to be able to call a given
      construction "a thesaurus"?
   c/ which elements and factors influence the organization
      of a thesaurus?
   d/ how should the degree of complexity and the number of
      information items contained in a thesaurus be evaluated?

2. I s   t h e   c o n c e p t   o f   a   t h e s a u r u s
   s u f f i c i e n t l y   c o m p l e t e   a n d   u n i -
   v o c a l l y   a n d   e x h a u s t i v e l y   d e f i n -
   e d,   o r   d o e s   i t   n e c e s s i t a t e
   f u r t h e r   a n a l y s i s?

3. W h a t   i s   t h e   r o l e   o f   a   t h e s a u r u s?
   a/ direct use in information and retrieval systems
   b/ in development of scientific information

4. H o w   c a n   t h e s a u r i   b e   c o n s t r u c t e d?
   a/ methods of compiling thesauri
   b/ possibilities and advantages of automation in compilation
      of thesauri

5. H o w   c a n   t h e s a u r i   b e   c l a s s i f i e d?
   According to:
   a/ branches /subject/
   b/ accuracy of definition of concept relations
   c/ the degree of hierarchy

d/ adopted types of basic components
e/ coding methods?

6. How can we define a descriptor?
   /Conditions which a key-word must fulfill in order to be-
   come a descriptor - the methods of constructing descrip-
   tors/.

7. What conditions must descrip-
   tors and thesauri fulfil in
   order to ensure their inter-
   branch and inter-language cor-
   relation?

8. Whet conditions must be fulfill-
   ed by descriptors and thesa-
   ·uri as tools for the further
   development of information?

   We consider the above topics only as general outlines,and
would be grateful if you widen the scope of the subject matter.
If in your opinion it is not necessary to discuss all the
items, please give your opinion only on such points which you
consider to be of utmost importance.

CONTENTS

- 186 -

1˙1

Annexes to the article:

J. Lojžišek, Statistical Analysis

of Documentation Files-SADF see page: 59-62

INPUT DATA FOR THE COMPUTER CONSISTING OF INDIVIDUAL ROO's
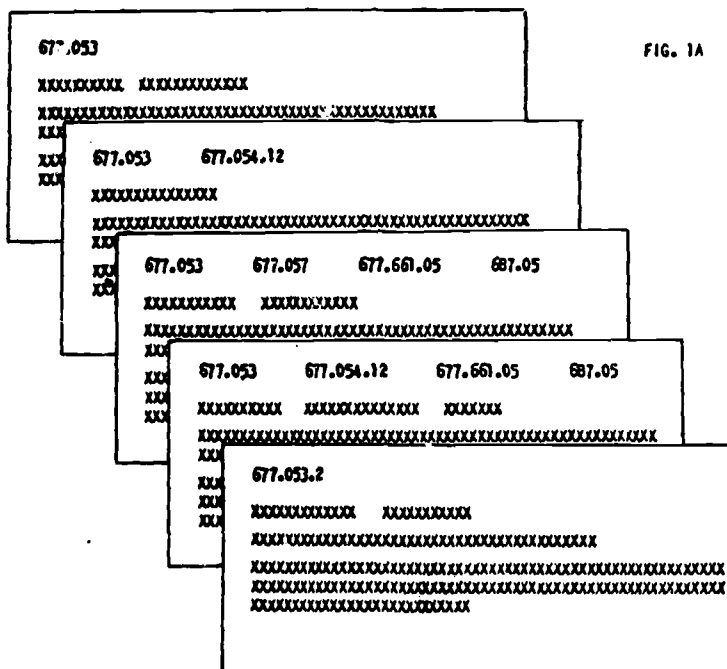
FIG. 1A

FIG. 1B

TABLE OF SINGLE INDEXING TERMS (SIT's)

SIMPLE INDEXING TERM   FREX.   RANK

.
.
.
.

621.335	1	345
621.335:2	1	346
621.335-2.024	1	347
621.335.025	1	348
621.335.2	100	6
621.335.2.024	23	96
621.335.2.024/.025	56	29
621.335.2.025	87	11
621.335.3.025	1	349
621.335.42	153	2
621.335.42-183.4	1	350
621.335.42-835	10	174
621.335.42.024	47	40
621.335.42.024/.025	11	162
621.335.42.025	67	23
621.336	5	244
621.355	2	311
621.355.2	1	351
621.355.8	3	292
621.39	2	312

.
.
.
.

SADF-2   FIG. 2

SURVEY OF SINGLE INDEXING TERMS (SIT's) FUNCTIONING

INDEX-2   FIG-3

1 - RANK
2 - FREQUENCY OF THE SIT
3 - PERCENTAGE OF THE DOCUMENTS
4 - PERCENTAGE OF SIT TOKENS
5 - ACCUMULATIVE PERCENTAGE OF SIT TOKENS
6 - ENTROPY INCREMENT
7 - ACCUMULATIVE ENTROPY
8 - RANK OF RDD
9 - FREQUENCY OF RDD
10 - PERCENTAGE OUT OF THE FREQUENCY OF SIT
11 - RANK OF CONCURRENT SIT
12 - FREQUENCY OF THE PAIR OF SIT's
13 - FREQUENCY OF THE PAIR OF SIT's IN PER CENT
14 - NUMBER OF RDD's CONTAINING THE PAIR OF SIT's

SURVEY OF SIT FUNCTIONS

RANK

SINGLE INDEXING TERM

	FREQV. JPZ	PROC. DOK.	PROC. JPZ	KUM.PROC. JPZ	ENTROPIE PAIR. CELK.		RANK RPZ	FREQV. RPZ	PROC. K.JPZ	RANK KOMB JPZ	FREQV. DVOJICE JPZ	PROC. DVOJICE K.JPZ	POCET RPZ S DVOJICI
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
① 665-42 161	3,4	2,2	2,2	,114	,1141		445 470 479	- - -	0,6 0,6 0,5	2 5 9	3 5 10	1,9 3,1 6,2	2 3 6
② 621,135-42 153	3,2	2,1	4,4	,110	,2242		447 447 486	- - -	0,7 0,7 0,7	1 5 6	3 13	2,0 0,7 0,5	2 11
③ 615,264-615,6 136	2,8	1,9	6,3	,101	,3254		536 558 575	- - -	0,7 0,7 0,7	4 5 6	11 4 4	8,1 2,9 2,9	6 1 2

1 - RANK
2 - FREQUENCY OF THE SIT
3 - PERCENTAGE OF DOCUMENT
4 - PERCENTAGE OF SIT TOKENS
5 - ACCUMULATIVE PERCENTAGE OF SIT TOKENS

6 - ENTROPY INCREMENT
7 - ACCUMULATIVE ENTROPY
8 - RANK OF RDD
9 - FREQUENCY OF RDD
10 - PERCENTAGE OUT OF THE FREQUENCY OF SIT

11 - RANK OF CONSUMER SIT
12 - FREQUENCY OF THE PAIR OF SIT's
13 - FREQUENCY OF THE PAIR OF SIT's IN PER CENT
14 - NUMBER OF RDD's CONTAINING THE PAIR OF SIT's

SKIP-2   FIG. 4

PRIMERNY POCET KP2 S JEDNIM JPZ   8,35033   a

PRIMERNY POCET KOMB. JPZ   10,00000   b

a - AVERAGE NUMBER OF RDD's CONTAINING A COMMON SIT
b - AVERAGE NUMBER OF CONCURRENT SIT's

SADF-2   FIG. 5

177

FREQUENCY DICTIONARY OF RETRIEVAL DOCUMENT DESCRIPTIONS (ROO's)

RANK	FREKV.	KUMUL. FREKV.	RELAT. F-EKV.	R.KUM. F-EKV.	PRIRUSTEK ENTROPIE	KUMULAT. ENTROPIE					
1	2	3	4	5	6	7	8	9	10	11	12

KOMPLEXNI POKACACI ZNAK

625.2-597.3:678.5   625.2-592.117

| 803 | 1 | 3886 | 0,0 | 81,2 | ,0018 | 4,88595 | 92 | 67 | 0 | 0 | 0 |

621.313.12/.13   621.335.2.025

| 304 | 1 | 3887 | 0,0 | 81,2 | ,0018 | 4,88772 | 93 | 11 | 0 | 0 | 0 |

621.313.12/.13   621.436   625.2.011,5   625.282-833.6

| 805 | 1 | 3888 | 0,0 | 81,2 | ,0018 | 4,88949 | 93 | 50 | 28 | 3 | ⊙ |

621.313.12/.13   625.2-5   625.28:621.317   621.335-42.025

| 806 | 1 | 3889 | 0,0 | 81,3 | ,0018 | 4,89126 | 93 | 57 | 56 | 23 | 0 |

621.313.12/.13   625.39

| 807 | 1 | 3890 | 0,0 | 81,3 | ,0018 | 4,89303 | 93 | 65 | 0 | 0 | 0 |

where: a/ RETRIEVAL DOCUMENT DESCRIPTION

1 - RANK
2 - FREQUENCY
3 - ACCUMULATIVE FREQUENCY
4 - RELATIVE FREQUENCY
5 - RELATIVE ACCUMULATIVE FREQUENCY
6 - ENTROPY INCREMENT
7 - ACCUMULATIVE ENTROPY
8 - 12 - RANKS OF THE SIT's IN THE RDO

SAUF-2   Pixu. b

PRIMERNY POCET DOKUMENTU OZNACENY JEDNIM KPZ   2,81033   a

PRIMERNE PROCENTO DOKUMENTU OZNACENYCH JEDNIM KPZ   ,058720   b

PRIMERNY PRIRUSTEK ENTROPIE   ,003805   c

a - AVERAGE NUMBER OF DOCUMENTS INDEXED WITH AN EQUIVALENT   RDD

b - AVERAGE PERCENTAGE OF DOCUMENTS INDEXED WITH AN EQUIVALENT   RDD

c - AVERAGE ENTROPY INCREMENT

SADF-2   FIG. 7

—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—

a   POCET VSECH JPZ   7184

b   POCET DOKUMENTU   4786

c   POCET CAST. SETRIDENYCH JPZ   1781

—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—x—xxx

a.   NUMBER OF ALL SIT OCCURENCES

b   NUMBER OF DOCUMENTS

c   NUMBER OF PARTIALLY ORDERED SIT's   (to be used by the com-
                                             pute operator)

SADF-2   FIG. 8

DIAGRAM OF THE RELATION BETWEEN THE RANK OF A RDD, ITS FREQUENCY AND THE ACCUMULATIVE VALUE OF THE RDD FREQUENCY

FREQUENCY OF RDD.

ACCUMULATIVE VALUE OF THE FREQUENCY OF RDD

RANK ("X" AXIS — LOGARITHMIC) ⟶

PER CENT ("Y" AXIS — LINEAR) ⟶

SADF-2.   FIG. 9

DIAGRAM OF THE RELATION BETWEEN THE RANK OF A RDD, ITS FREQUENCY AND THE ACCUMULATIVE OF THE RDD FREQUENCY

ACCUMULATIVE VALUE OF THE FREQUENCY OF RDD

FREQUENCY OF RDD

RANK ("X" AXIS — LINEAR)

PER CENT ("Y" AXIS — LINEAR)

SADF-2 FIG. 10

439
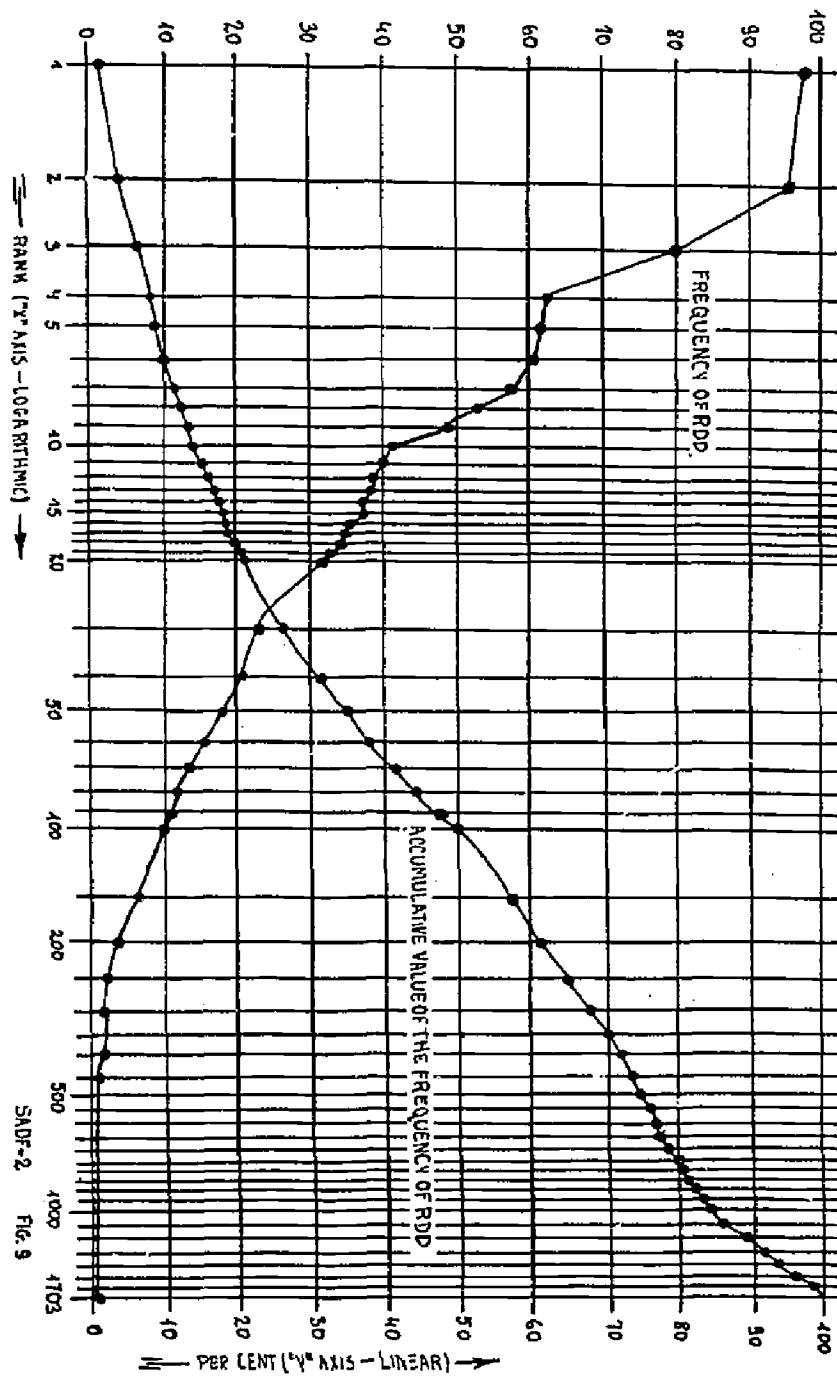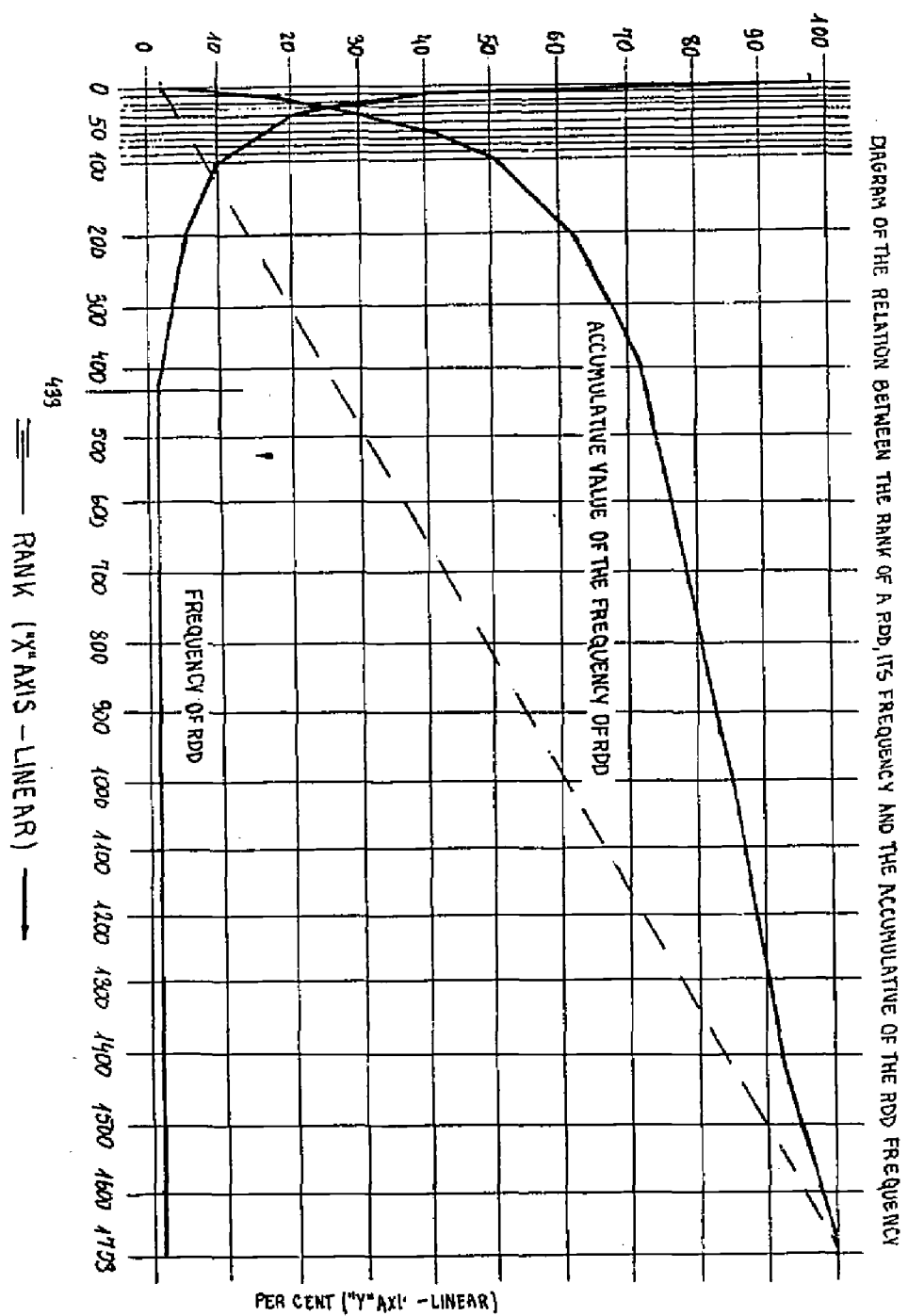
R	f	R.k.f.
1	98	2,0
2	96	4,1
3	80	5,6
4	63	7,0
5	62	8,3
6	61	9,6
7	57	10,8
8	53	11,9
9	49	12,9
10	42	13,8
11	40	14,6
12	39	15,5
13	39	16,3
14	38	17,1
15	38	17,9
16	37	18,6
17	36	19,4
18	36	20,1
19	33	20,8
20	32	21,5

R	f	R.k.f.
30	24	27,1
40	21	31,7
50	18	35,8
60	16	39,2
70	14	42,2
80	12	44,8
90	11	47,2
100	10	49,4
150	6	57,3
200	4	62,1
250	3	65,4
300	2	67,9
350	2	70,0
400	2	72,1
450	1	73,8
500	1	74,9
550	1	75,9
600	1	77,0
650	1	78,0
700	1	79,0

R	f	R.k.f.
750	1	80,1
800	1	81,1
850	1	82,2
900	1	83,2
950	1	84,3
1000	1	85,3
1200	1	87,4
1200	1	89,5
1300	1	91,6
1400	1	93,7
1500	1	95,8
1600	1	97,6
1703	1	

R - RANK

f - FREQUENCY

R.A.f. - RELATIVE ACCUMULATIVE FREQUENCY

RELATIVE ACCUMULATIVE FREQUENCY REACHES 50 PER CENT WITH THE RANK OF 103

FREQUENCY f = 2 BEGINS WITH THE RANK OF 274

f = 1 BEGINS WITH THE RANK OF 433

SADF-2   FIG. 11

A	B	C	d1	d2	d3	d4	d5	e1	e2	e3	e4	e5
1-1	98	1	2					2,0				
2-2	96	1	1					4,1				
3-3	80	1	1					5,6				
4-4	63	1	1					7,0				
5-5	62	1	1					7,3				
6-6	61	1	1					9,6				
7-7	57	1	1					10,8				
8-8	53	1	1					11,9				
9-9	49	1	1					12,9				
10-10	42	1	1					13,8				
11-11	40	1	1					14,6				
12-13	39	2	2					16,1				
14-15	38	2	2					17,9				
16-16	37	1	1					18,6				
17-18	36	2	2					20,1				
19-19	33	1	1					20,8				
20-21	32	2	2					22,2				
22-22	30	1	1					22,8				
23-23	29	1	1					23,4				
24-24	27	1	1					24,0				
25-26	26	2	2					25,1				
27-27	25	1	1					25,6				
28-31	24	4	4					27,6				
32-34	23	3	3					29,0				
35-37	22	3	3					30,4				
38-41	21	4	4					32,2				
42-45	20	4	4					33,8				
46-48	19	3	3					35,0				
49-50	18	2	2					35,8				
51-55	17	5	5					37,5				
56-60	16	5	5					39,4				
61-62	15	2	2					39,8				
63-71	14	9	8	1				42,2	42,5			
72-75	13	4	4					43,3	43,6			
76-85	12	10	9	1				45,5	46,1			
86-92	11	7	7					47,1	47,7			
93-104	10	12	10	2				49,2	50,2			
105-113	9	9	9					50,9	51,9			
114-124	8	11	11					52,7	53,7			
125-137	7	13	13					54,6	55,6			
138-156	6	19	16	3				56,7	58,0			
157-175	5	19	14	5				58,2	60,0			
176-207	4	32	17	14		1		59,4	62,6		62,7	
208-273	3	66	32	34				60,1	66,7		66,8	
274-432	2	159	37	110	11	1		61,8	72,8	73,3	73,4	
433-1703	1	1271	95	687	349	77	63	63,8	89,1	97,0	98,7	100
Σ =	46	1703	344	857	360	79	63					

where: A - IDENTIFICATION OF ISO-FREQUENCY GROUP

B - FREQUENCY

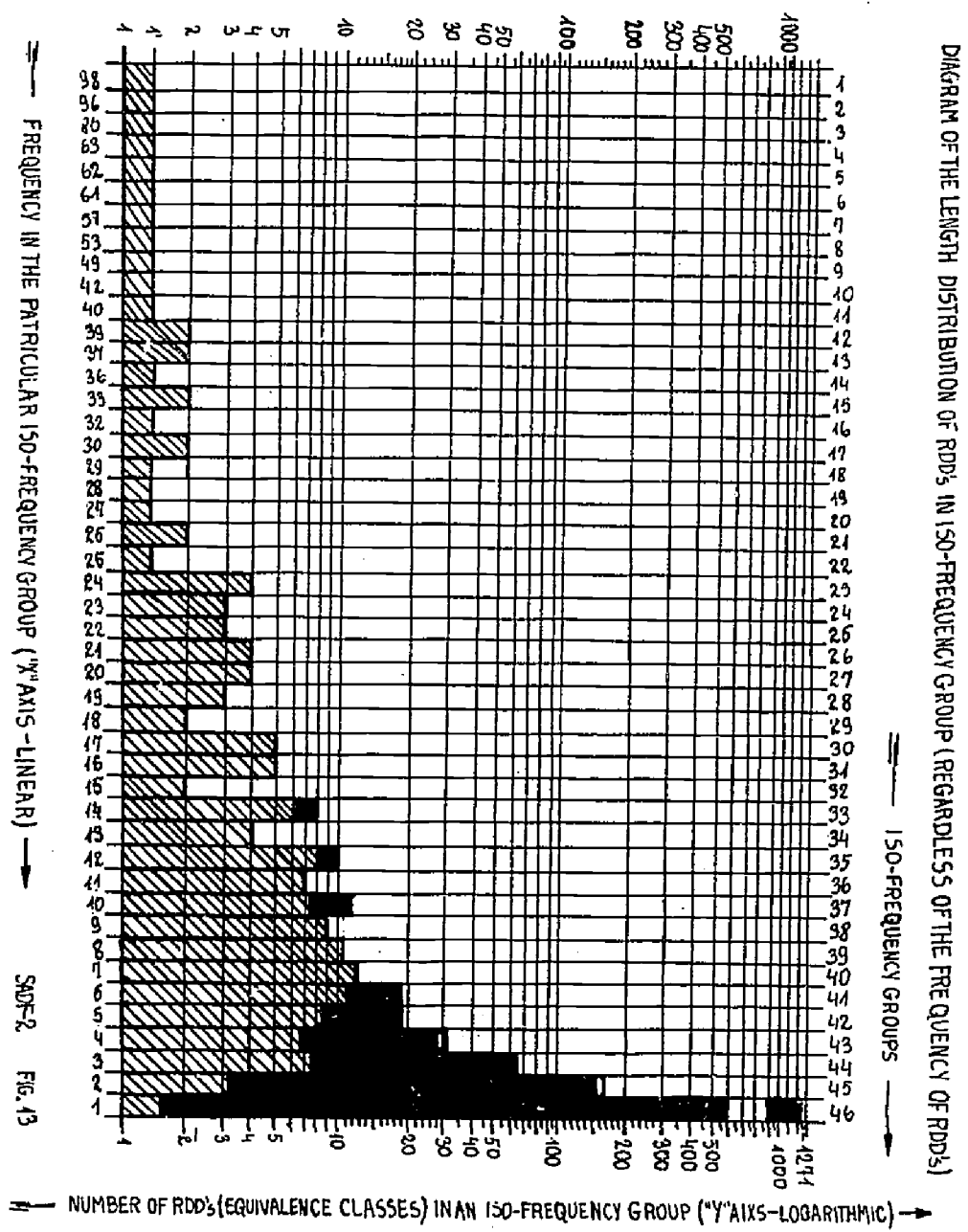C - NUMBER OF ROO's IN THE ISO-FREQUENCY GROUP

D - NUMBER OF ROO's OF THE LENGTH OF :

d_1 - 1 SIT, d_2 - 2 SIT's, d_3 - 3 SIT's, d_4 - 4 SIT's, d_5 - 5 SIT's

E - PERCENTAGE OF THE ROO OCCURENCES (TOKENS - REGARDING THE FREQUENCY OF THE ROO's) OF THE LENGTH OF :

e_1 - 1 SIT, e_2 - 2 SIT's, e_3 - 3 SIT's, e_4 - 4 SIT's, e_5 - 5 SIT's

SADF-2   FIG. 32

DIAGRAM OF THE LENGTH DISTRIBUTION OF RDD's IN ISO-FREQUENCY GROUP (REGARDLESS OF THE FREQUENCY OF RDD's)

FREQUENCY IN THE PATRICULAR ISO-FREQUENCY GROUP ("Y"AXIS-LINEAR) →

SIDE=2.   FIG. 13

NUMBER OF RDD's (EQUIVALENCE CLASSES) IN AN ISO-FREQUENCY GROUP ("Y"AIXS-LOGARITHMIC) →

ISO-FREQUENCY GROUPS →

PREQUENCY DICTIONARY OF RETRIEVAL DOCUMENT DESCRIPTIONS /RDD´s/

SADP-2   FIG. 14

RETRIEVAL DOCUMENT DESCRIPTION

RANK	FREKV.	KUMUL. FREKV.	RELAT. FREKV.	R.KUM. FREKV.	PRIRUSTEK ENTROPIE	KUMULAT. ENTROPIE	RANKY JPZ v KPZ				
	2	3	4	5	6	7	8	9	10	11	12
1											
625.232 30	24	1296	0,5	27,1	,0266	1,24976	45	0	0	0	0
625.243.5 64 ⑭	625.241	1935	0,3	40,4	,0171	1,98792	53 16	0	0	0	0
625.2.9 2.813 ⑬		2059	0,3	43,0	,0160	2,13949	II	0	0	0	0
539.3/.4 76 ⑫	621.791	2097	0,3	43,8	,0150	2,18660	20 24	0	0	0	0
625.245.9 1154	625.2427 ①	625.2-592.5 4237	0,0	625.2-592.001 88,5	625.2-597 ,0018	5,50738	169 141	109	24	44	

1 - RANK
2 - FREQUENCY
3 - ACCUMULATIVE FREQUENCY
4 - RELATIVE FREQUENCY
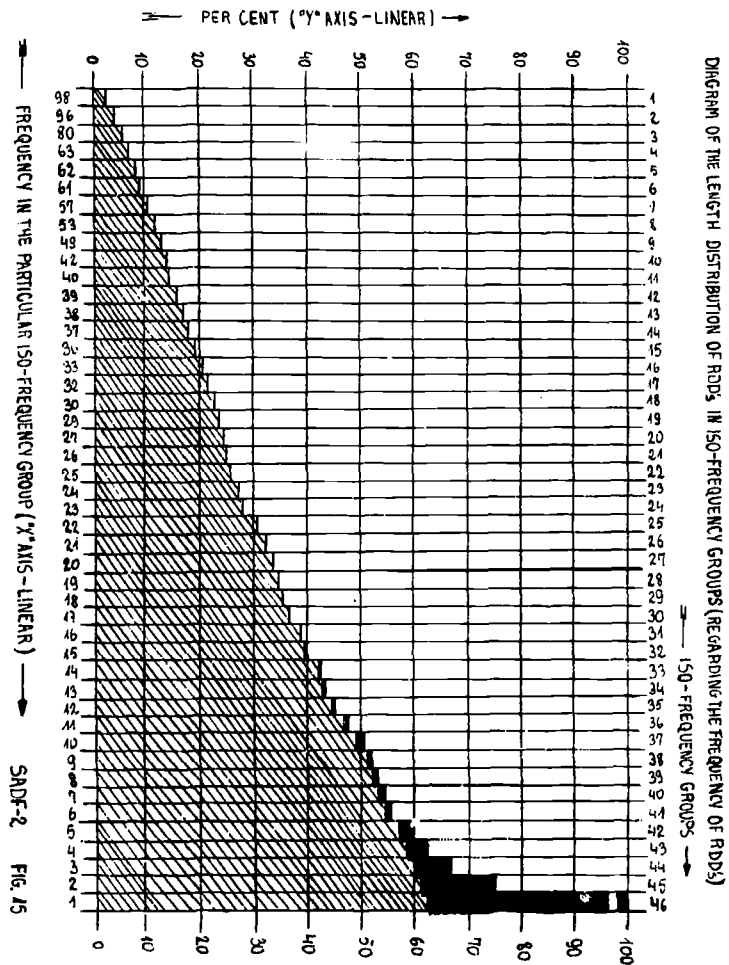5 - RELATIVE ACCUMULATIVE FREQUENCY
6 - ENTROPY INCREMENT
7 - ACCUMULATIVE ENTROPY
8 - 12 - RANKS OF THE SIT´s IN THE RDD

625.2-592.001

625.2-597

DIAGRAM OF THE LENGTH DISTRIBUTION OF RDDs IN ISO-FREQUENCY GROUPS (REGARDING THE FREQUENCY OF RDDs)

PER CENT ("Y" AXIS — LINEAR)

ISO-FREQUENCY GROUPS

FREQUENCY IN THE PARTICULAR ISO-FREQUENCY GROUP ("X" AXIS — LINEAR)

SADF-2   FIG. 15

SURVEY OF SINGLE INDEXING TERMS (SIT's) FUNCTIONING

SINGLE INDEXING TERM

RANK	FREKV. JPZ	PROC. DOK.	PROC. JPZ	KUM.PROC. JPZ	ENTROPIE PAIR. CELK.		RANK KPZ	FREKV. KPZ	PROC. K JPZ	RANK KOMB JPZ	FREKV. DVOJICE JPZ	PROC. DVOJICE K JPZ	POCET KPZ S DVOJICI
1	2	3	4	5	6	7	8	9	10	11	12	13	14
⑬ 225.245.53													
	32	0,7	0,4	60,9	,033	3,912	673	1	3,1	6	1	3,1	1
							674	1	3,1	7	3	9,4	2
							675	1	3,1	10	4	12,5	4
							676	1	3,1	16	3	9,4	3
							811	1	3,1	20	4	12,5	3
							1005	1	3,1	25	3	9,4	3
							1045	1	3,1	27	1	3,1	1
							1153	1	3,1	104	1	3,1	1
							1197	1	3,1	134	1	3,1	1
							1206	1	3,1	144	1	3,1	1
							1433	1	3,1	169	1	3,1	1
							1590	1	3,1	179	1	3,1	1
							1647	1	3,1	181	1	3,1	1
							324	2	6,3	258	2	3,1	1
							325	2	6,3	336	1	3,1	1
							62	15	46,9	395			

1 - RANK
2 - FREQUENCY OF THE SIT
3 - PERCENTAGE OF THE DOCUMENTS
4 - PERCENTAGE OF SIT TOKENS
5 - ACCUMULATIVE PERCENTAGE OF SIT TOKENS

6 - ENTROPY INCREMENT
7 - ACCUMULATIVE ENTROPY
8 - RANK OF RDD
9 - FREQUENCY OF RDD
10 - PERCENTAGE OUT OF THE FREQUENCY OF SIT

11 - RANK OF CONCURRENT SIT
12 - FREQUENCY OF THE PAIR OF SIT's
13 - PREQUENCY OF THE PAIR OF SIT's IN PER CENT
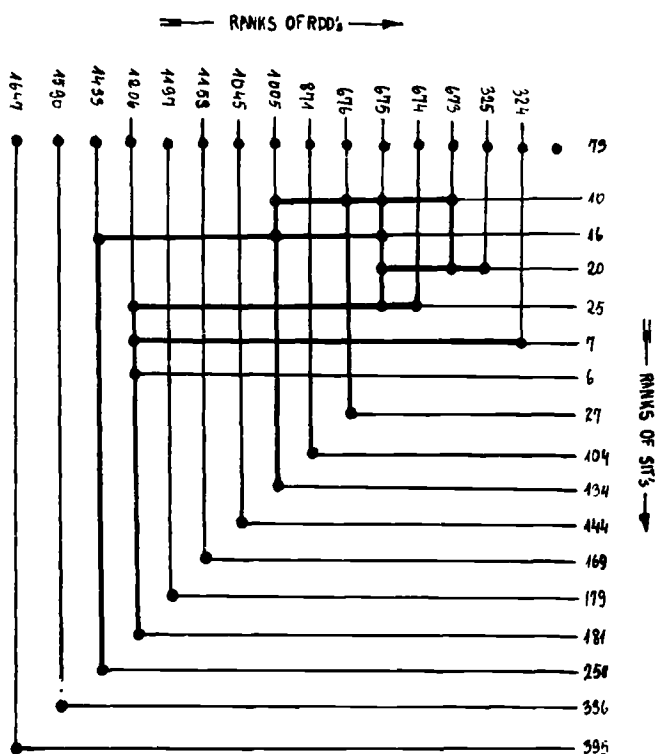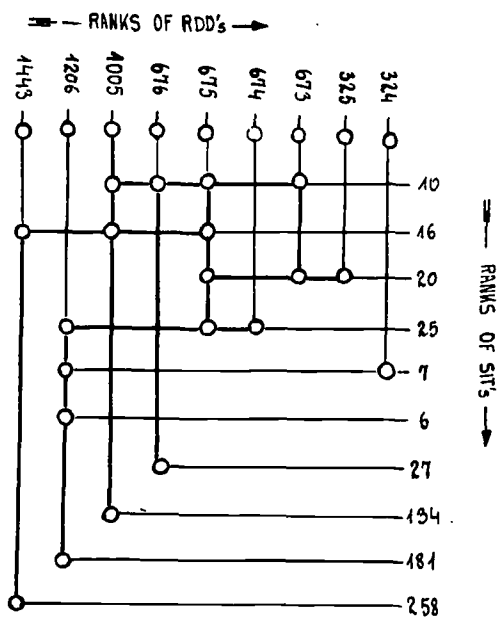14 - NUMBER OF RDD's CONTAINING THE PAIR OF SIT's

SADT-2

FIG. 21

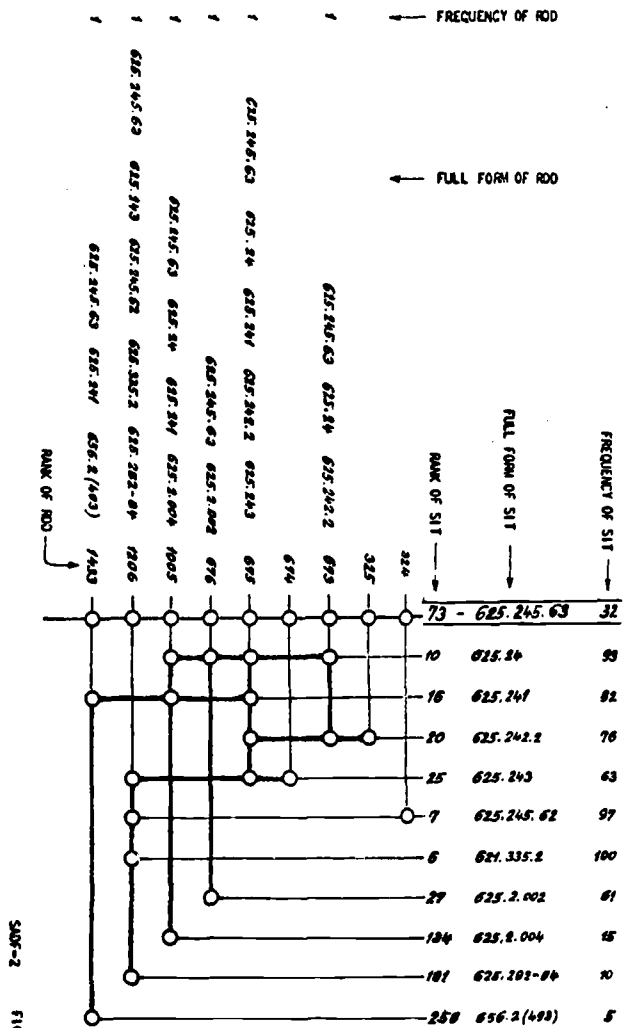DIAGRAM OF THE COMBINATION POWER OF THE SIT OF THE RANK OF 73

SADF-2.   FIG.22

ABBREVIATED DIAGRAM OF THE COMBINATION POWER OF THE ST OF THE RANK OF T3

RANKS OF RDD's →

RANKS OF SIT's →

SADF-2   FIG. 2.3

ABBREVIATED DIAGRAM OF THE COMBINATION POWER OF THE SIT OF THE RANK OF 73 (WITH FULL FORM OF SIT's AND RDO's)

FIG. 24

SAID-2

191