Hemalatha Iyer
School of Information Science and Policy,
SUNY, Albany, New York

Natural Language Representation: Transformational Rules

Iyer, Hemalatha: **Natural language representation: transformational rules.**
Iut.Classif. 17(1990)No.1, p.8-13, refs.

This paper presents a comparison of facet structure and linguistic structure of subject prepositions, resulting in standardisation of prepositions connecting facets/facet speciator. The facet structure is derived using the general theory of classification developed in India, and for the linguistic analysis Halliday's system is used. A set of transformational rules for switching from facet structure to the natural language representation is also presented.

(Author)

1. Introduction

A subject heading or a subject proposition in a pre-coordinated indexing system is made up of several constituent elements. A semantic and syntactic relational structure can be attributed to it. Such a relational structure may be generalized into a postulate. The general theory of classification developed in India, based on Dr.S.R.Ranganathan's postulates and principles provides a kind of typology of generic relations resulting in a facet structure, which can be used for generating an organized set of subject propositions.

2. Rangauathan's Five Fundamental Categories

Ranganathan's approach to structuring of subjects is based on a postulational approach. It centers around the concepts of Basic Subject (BS); the five fundamental categories (FC), Personality (P), Matter (M), Energy (E), Space (S), and Time (T) (1). Personality is the core entity of a subject statement. Ranganathan considered it the most ineffable one for definition (6). For recognition of *Personality*, he suggested the Method of Residues. However, this was not found adequate. As Gopinath writes:

> The problems in the recognition of the (FC) Personality is not definitional, but contextual. The semantic and syntactic aspects in the formation of these compound subjects and the generalization of these structures to a modal base ... that is, a basic subject - - sets the difficulties in the recognition of Personality (2).

Gopinath has analyzed the problems in identification of fundamental categories in interdisciplinary subjects and has framed criteria and methods for the same.

Matter. Matter connotes a property or materialness of the focal idea of the subject statement. The material constituent of the focal idea was considered to be Matter. The recognition of a qualifier concept in 1963-64 led to the recognition of the material constituent as the qualifier. Such qualifiers are known as speciators or (SP) (5). The property ideas were deemed to be manifestations of (FC) matter. In this study (M) denotes a manifestation of the property type concepts.

Energy. Energy connotes an action in relation to the focal idea. The concept of Energy was constantly being examined to clarify its connotation. In 1952 'Energy' was defined to include 'Problem, Action, Quality', etc. (3). Ranganathan stated:

Energy manifests itself either as motion, interaction or mutual action of some kind or as one of the isolates postulated to be 'Energy', such as those denoted by the terms 'Morphology', 'Physiology', 'Disease', 'Ecology', 'Phylogeny, 'Ontogamy' and their equivalent. (4)

Space. The concept of (FC) 'Space' is in accordance with what is commonly understood by that term -- the surface of the earth, the space inside and outside it. The geographical areas and physiographic features are manifestatitons of (FC) Space.

Time. This connotes the usual time concepts such as milenium, century, decade, year, etc.

The five fundamental categories are interrelated, and with this in view, Ranganathan sequenced them as "PM EST", in the order of decreasing concreteness of categories. J.M.Perreault, while suggesting a scheme of categories and relators to be used with the Universal Decimal Classification, comments that the Colon Classification is most satisfactory in terms of the syntax that permits sublety (7).

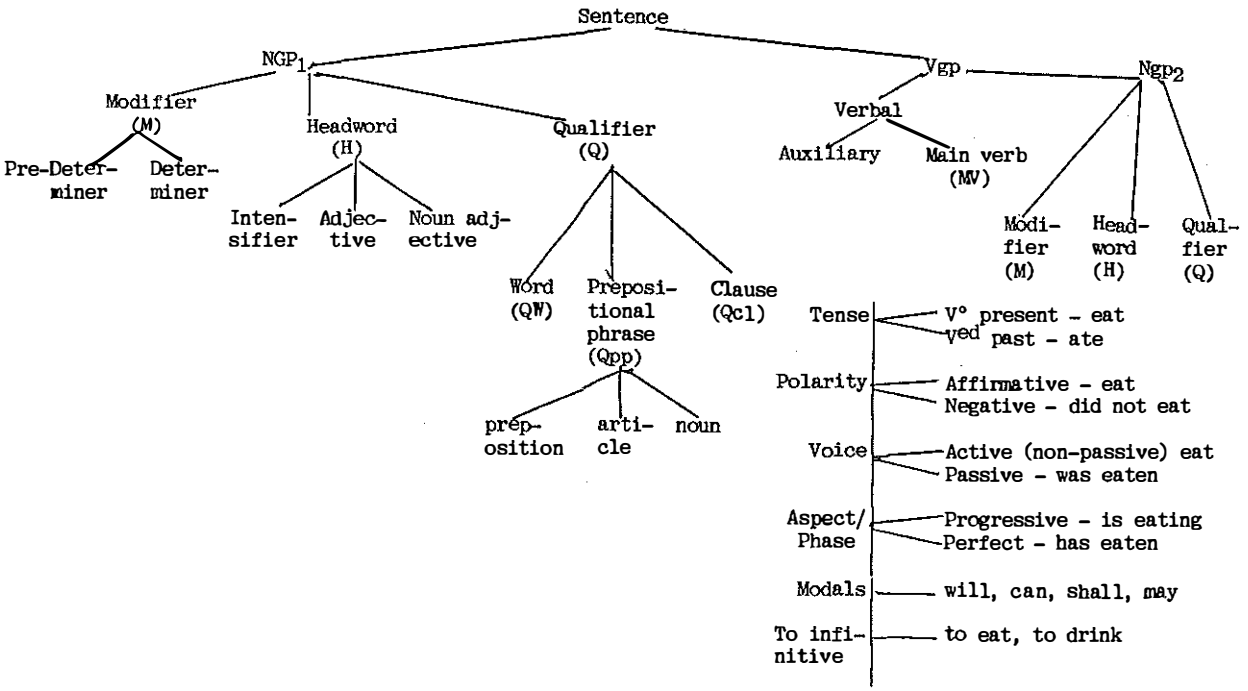
The facet structure of a subject proposition can be correlated to similar structures in linguistics; in particular, there is a parallel in the inter-constituent structure of a formal language in Halliday's System and Structures.

This our study aims to make a comparison between linguistic structure and facet structure and to formulate rules for transformation from facet structure to natural language representation.

3. Halliday's System

Halliday's inter-constituent analysis recognizes the noun group and the verb group. The noun group consists of a Modifier (M), the Headword (H) and the Qualifier

Diagram 1 Halliday's Interconstituent Analysis of entence



(Q). The Modifier consists of the following constituents: 1) Pre-determiner, 2) Determiner, 3) Numeral, 4) Intensifier, 5) Adjective, and 6) Noun adjective. The determiner, in turn, can be an article (a, an, the), a demonstrative or a possessive case (e.g. John's). A numeral can be an ordinal (1st, 2nd) or a cardinal (1, 2, 3, 4). Intensifiers are words that refer to 'degree' of possession of some attribute, for example: very, few, etc. A noun adjective is a noun acting as an adjective, for example: 'The Madras team'. Here 'Madras' is a proper noun, acting as an adjective to the noun 'team'.

Any or all of the above mentionend modifiers can modify the Headword. The Headword is the most important word in the phrase. It is invariably a noun. The Qualifiers are those words that qualify the Headword and usually follow it. The qualifier again can be of three types. A word (QE), a prepositional phrase (QPP) or a clause (QcI). A prepositional phrase in a Qualifier consists of the preposition (in, at, on, etc.), an article and a noun. The Qcl takes 'who' or 'which' as subordinate clause.

The verb group consists of a verbal and a noun group. The verbal group again consists of the auxiliary and the main verb (MV). Auxiliary presents a system of choices, like Tense, Polarity and Voice, and other elements like Aspect, Phase, Modal and to-infinitive (8).

Representing the interconstituent analysis in a diagrammatic form gives rise to the structure as given in Diagram 1.

For the purpose of this study a random sample of 100 titles listed under the heading 'Urban Sociology' in the 1980 Sociological Abstracts was taken. These titles were analyzed for the phrase structure using Halliday's system, and facet analyzed using Ranganathan's principles and postulates and the Colon Classification, Ed.7. Urban Sociology is an environmented Basic Subject, i.e. the basic subject 'Sociology' is qualified by the environment 'Urban region'. Since it is still sociology, the concept of community remains as the (P) facet. (M) is derived from two parent basic subjects 'Town planning' as well as 'Sociology', e.g. 'layout', and 'neighborhood', (E) denotes some kind of action, either a common action isolate or a special action isolate. It pertains to both social work and urban planning. A concept can be in a static state or a dynamic state, e.g. 'change' in a static state refers to a change that has occurred, i.e. the changed state of the thing, therefore (M). The dynamic state refers to the 'process of change', hence it becomes an action performed on the thing, therefore (E).

The common isolates of the Colon Classification, 7th edition, are also used for analysis. Thus the component facets of urban sociology involve the following types of concepts:

Sociology - by environment (Sp to BS) i.e. (Sociology - urban)
Community - (1P1) by urban environemt (Sp) to (1P1)
Property - (M)
Energy - (E)
Agent of action (Sp) to (E), Method of Action (Sp) to (E).

Example:

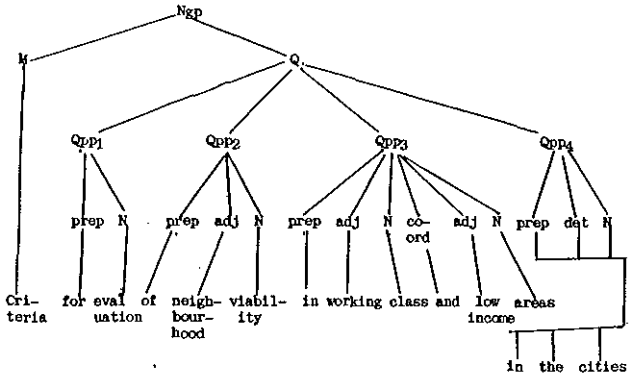
Title: Criteria for evaluation of neighborhood viability in working class and low income areas in the cities.

Facet Structure Analysis: Urban Sociology (BS), city (P), working class and low income area (P2); neighborhood viability (M); criteria (M2).

4. Phrase Structure Analysis of the Titles

The phrase structure analysis of the titles indicated that 71% of the titles are composed of one noun group.

Phrase Structure Analysis



Within this set, in 75% of the titles there is an occurrence of a modifier. The modifier consists of one or more adjectives. The incidence of Qpp ranges from one to four, though the majority of titles contain only one Qpp. The Qpp's contains a preposition followed by one or more adjectives and a noun.

25% of the titles are composed of two noun groups, 2% of the titles consist of three noun groups and the remaining 2% of the Qpp's only.

From the patterns observed, in the field of urban sociology, the community facet is always present. In most of the cases the concept is associated with property, which may be the behavioral or the environmental aspect of the community.

Thus we may see that the compound subjects associated with Urban Sociology highlight the incidence of PME/PM more than any other relational structure. The incidence of (S) and (T) is relatively low.

Table 1 Typology of Relational Structures and their Incidence

S. No.	Faceted Structure of Titles	Frequency of Incidence
1	BS, P "ACI	1
2	BS, P:E	2
3	BS, P:E "ACI	1
4	BS, P:E 'T	1
5	BS, P:E 'T	1
6	BS, P:E. S'T	2
7	BS, P:E:E	1
8	BS, P:E; M	1
9	BS, P:E - Sp	2
10	BS, P:E - Sp; M	1
11	BS, P;M	10
12	BS, P;M 'T	2
13	BS, P;M: E "ACI	1
14	BS, P;M: E	9
15	BS, P;M: E 'T	1
16	BS, P;M: E: 2E	1
17	BS, P;M: E: 2E - Sp "ACI	1
18	BS, P;M: E - Sp	9
19	BS, P;M: E - Sp.S	1
20	BS, P;M: E - Sp - Sp	1
21	BS, P;M; M_2	6
22	BS, P;M; M_2 'T	1
23	BS, P;M - Sp: E	1
24	BS, P;M - Sp: E: 2E	1
25	BS, P;M - Sp : E - Sp	1
26	BS, P;M - Sp; M_2	3
27	BS, P;M - Sp - Sp; M_2	1
28	BS, P,P_2	2
29	BS, P, P_2 S	1
30	BS, P, P_2: E 'T	2
31	BS,P, P_2: E:2E - Sp	1
32	BS,P, P_2; M	4
33	BS,P, P_2; M:E; 2M	1
34	BS,P, P_2 - Sp	1
35	BS,P - Sp	1
36	BS,P - Sp: E	3
37	BS,P - Sp:E	2
38	BS,P - Sp: E:2E	2
39	BS,P - Sp:E:2E - Sp	1
40	BS,P - Sp; M	6
41	BS,P - Sp - Sp; M	1
42	BS,P - Sp; M:E	2
43	BS,P - Sp; M; M_2 "ACI	1
44	BS,P - Sp; M; M_2	2
45	BS,P - Sp; M -Sp - Sp	1
46	BS,P - Sp, P_2	1
47	BS,P - Sp, P_2 -Sp	1
48	BS,P - Sp, P2 - Sp; M	1

Table 2

Facet Structure Order and the Relational Indicators

Title and the Order of Facets	Relational Indicator	Facet Structure Structure
Powerlessness in racially changing neighbourhood (Property, thing)	Thing, property- 'in' Property, property- 'in' Property, qualifier- adjective	US(BS), urban area (P); neighbourhood (M)- racially changing(Sp); powerlessness (M_2)
Residential dissatisfaction in the crowded urban neighbourhood (Property, thing)	Thing, property-' adjective' thing, property second -'in' Property Qualifier- 'adjective'	US(BS), urban area (P); neighbourhood-crowded(Sp); residential dissatisfaction (M_2)
Social and demographic characteristics of structural classification for construction of typology of city (Action, property, thing)	Action, property- 'of' Property, thing-'of' Action, Qualifier- 'for' Qualifier, qualifier- 'of'	US(BS), city(p); Typology(M): Construction (E) (of)- structional classification (Sp) (by) - social and demographic characteristics (SP)

5. Matching the Phrase Structure with the Facet Structure

The order of concepts in the facet structure differed from the order of the same concepts in the phrase structure in 71% of the titles. Out of this set, in 61% of the titles, the facet structure order was the reverse of the phrase structure order. 23% of the titles of the facet structure order matched with the phrase structure order. 6% of the titles had just one facet. Comparison revealed that in the majority of cases, the title involved a reversal of the facet structure.

In an effort to standardize the prepositions combining two given facets, giving rise to a different typology of relations, all titles were facet analyzed and the order of these facets in the respective titles was determined. Table 2 presents examples of such an anlysis.

From the foregoing analysis, a pattern of relational indicators was derived based on the typology of relations. (see table 3, p.11)

Prepositions dominate as relational indicators. The preposition 'in' occurs more frequently in thing-property relations. This accounts for 43 % of the occurrence of the preposition 'in'. It also acts as a connector between thing-action, thing-space, thing-viewpoint, action-space, action-time, thing-qualifier, property-levels. However, these occur less frequently. It may be observed that out of all the relational indicators used, 'in' accounts for 33%.

The relational indicator 'of' occurs as a connector between property-action, thing-action, and property-viewpoint. It has the maximum incidence as thing-property and property-action and thing-action relators. Apart from these, it also occurs in property-viewpoint. It may be observed that 'of' accounts for 36% of all cases.

The preposition 'for' occurs in the action-qualifier, thing-action relations.

The conjunction 'and' occurs in thing-property, thing-action type of relations. In the thing-property type of relation it occurs in 2% of the cases and in thing-action type of relation 3.8% of the cases.

The other relational indicators are 'within', 'which', 'through', 'between', 'with'. The incidence of these is low.

Table 3
Typology of Relation and Relational Indicator

Type of relation	Relational indicator	Frequency of Occurrance: Phrase Structure: reverse of facet structure	differs from facet structure	agrees with facet structure	Total
1. Thing, Property	of	7	1	11	19
	in	24	-	-	24
	within	1	-	-	1
	subtitle	1	-	-	1
	adjective	6	1	-	7
	and	2	1	-	3
	which	-	1	-	1
2. Action, Property	of	10	1	-	11
	in	3	-	-	3
	adjective	1	-	1	2
3. Thing, Action	of	14	-	-	14
	in	1	-	-	1
	adjective	1	1	-	2
	and	2	1	2	5
	through	1	-	-	1
	subtitle	1	-	-	1
4. Thing, Time	in	1	-	-	1
	between	1	-	1	2
5. Thing, Space	in	1	-	1	2
6. Thing, Time	in	1	-	1	2
7. Thing, View point	in	-	-	1	1
8. Thing, Case study	with	1	-	-	1
9. Property View Point	of	-	-	1	1
10. Space, Case study	adjective	-	-	1	1
11. Action, space	in	-	-	2	3
12. Action, time	in	-	-	1	1
13. Property, qualifier	adjective	5	-	2	7
14. Thing, qualifier	in	-	-	1	1
	adjective	7	-	1	8
15. Action, qualifier	for	1	-	-	1
16. Levels: Thing, thing	adjective	-	-	2	2
	of	1	-	1	2
	in	3	-	-	3
17. Property, property	in	1	-	-	1

In conclusion, the prepositions 'in' and 'of' dominate. The thing-property relations are connected mainly by the prepositions 'in' and 'of'. The thing-action and property-action relation is connected mainly by the preposition of.

In organizing information, the representation of the subject of the document, the facet structure orders the component concepts in the subject according to a predetermined sequence. This also facilitates browsing. However, the facet structured representation is not as effective as natural language in communicating the subjet of the document to the user.

Amongst several others, some of the most desirable characteristics so an indexing system from the searcher's point of view is, firstly, *clarity,* which refers to how likely it is that the index string will not be misinterpreted and to how readily a searcher can correctly grasp the meaning of the index string, or, in this case, the faceted representation. Secondly, *collocation*, which refers to placing similar index strings together and separating dissimilar ones. Thirdly, eliminability, which refers to how quickly a searcher can decide that the index string he is examining is irrelevant to him. This is partly dependent on clarity (9). Thus in an information sytem, it may be useful to provide for subject access using the faceted representation and also state the subject of the document in natural language. To be able to do so, the system should have the capability of switching from faceted representation to natural language representation. The following sections of this paper formulate the transformation rules to automate this process.

Examples:

1) Faceted representation (FR)
   Urban Sociology, Urban area; change: evaluation; equilibrium and adaptive approaches
   Natural language representation (NLR)
   Equilibrium and adaptive approaches in the evaluation of change in urban areas
2) FR
   History: classification
   NLR
   Classification of history
3) FR
   Urban sociology, city, working class and low income areas; neigborhood viability: evaluation; criteria
   NLR
   Criteria for evaluation of neighborhood viability of working class and low income areas in cities.

6. Transformational Rules

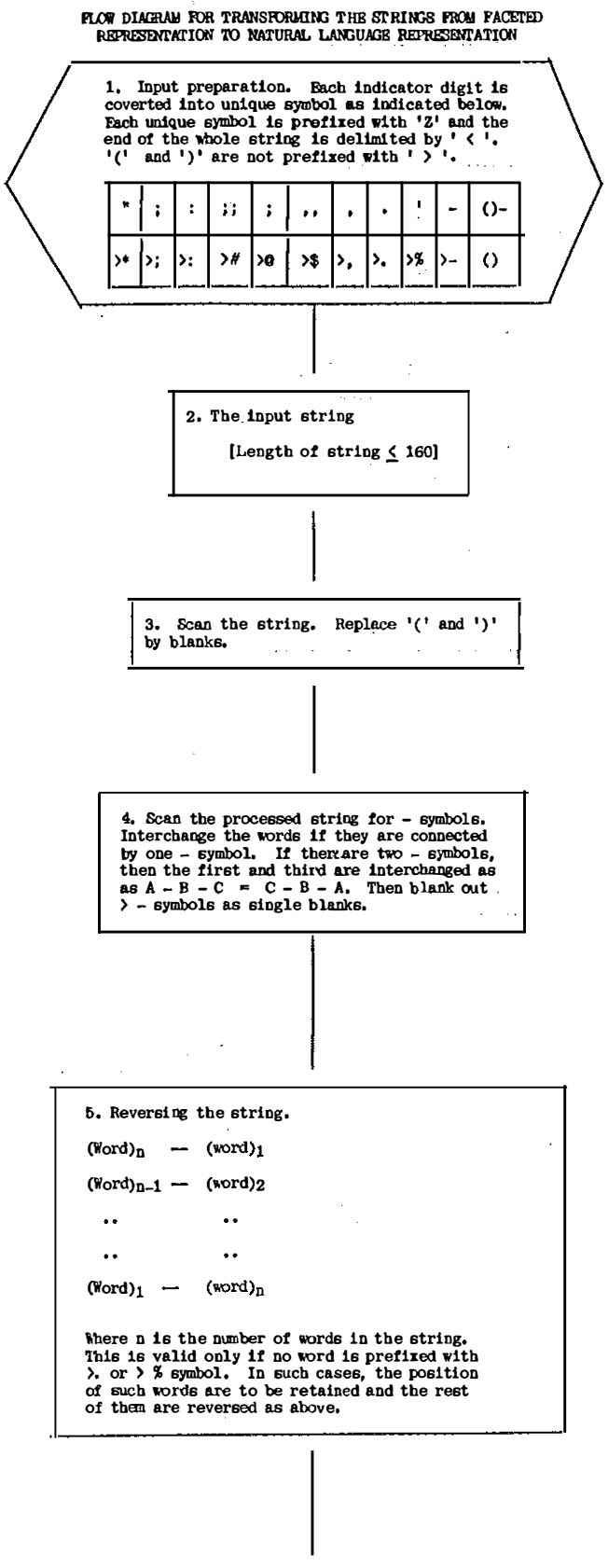
In the following the generalized facet formula for Urban Sociology is given derived on the basis of the analysis of the titles in the sample:

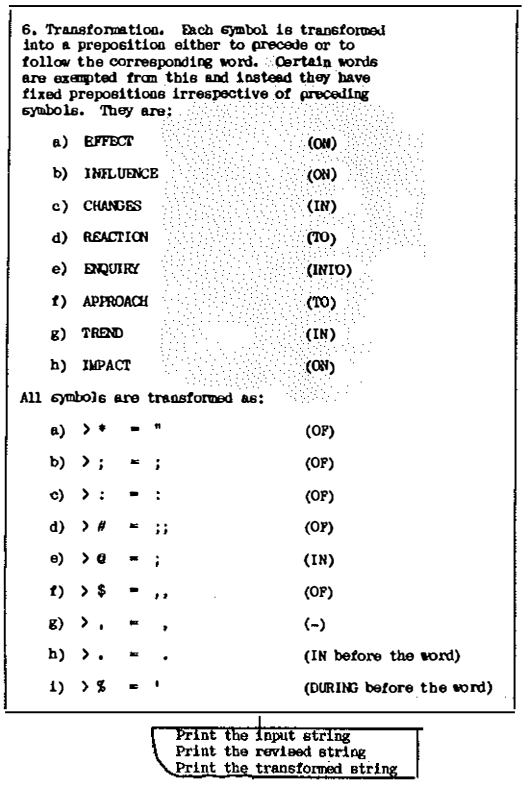
BS,P,,P2;M;;M2:E;2M:2E"ACI.S'T

In this formula the following connecting symbols for the facets have been used (new in CC Ed.7); they are adaptations of Ranganathan's 'connecting digits':

P is connected by ','
P2 is connected by ',,'
M and 2M are connected by ';'
M2 is connected by';;'
E is connected by ':'
ACI is connected by '"'
S is connected by '.' and
T is connected by "'

This program was executed on the sample of the said 100 titles. The transformation from the facet structure resulted in a meaningful natural language representation for altogether 997 titles.





7. Conclusion

The main focus of this study is standardization of relational indicators with reference to typology of facet relations; the correlation of the facet order with natural language; transformational rules for automated switching from the facet structure to natural language representation. Designing information systems with such a capability of automated switching would make them more effective.

The pre-coordinated index string would facilitate collocation and browsing, while the natural language representation would help the user to interpret the subject of the document accurately.

Further research needs to be done to test the rules for possibilities of operation in other subject areas and also in multilingual contexts.

References

(1) Ranganathan, S.R.: Library classification: Fundamental procedures. 1944
(2) Gopinath, M.A.: An analysis of the problem in the recognition of the manifestations of fundamental categories in the interdisciplinary subjects. Dharwar: Karnatak University. Ph.D.Dissertation 1980. p.198
(3) Ranganathan, S.R.: Colon. Classification. 5th ed. Bombay 1952.
(4) Ranganathan, S.R.: Colon. Classification 6th ed. Bombay 1961.
(5) Ranganathan, S.R.: Design of Depth Classification Methodology. Libr.Sci.Slant Doc. (1964)1, Paper A
(6) Ranganathan, S.R.: Prolegomena to library classification. Bombay: Asia Publ.House 3rd.ed 1967. Sec.RB7
(7) Perreault, J.M.: Towards a theory for UDC. London: C.Bingley 1969. p.152
(8) Kress, G.R. (Ed.): Halliday: System and function in language. 1976
(9) Craven, T.C.: String indexing. Orlando, Fla.: Academic Press 1986. XI,246p.

UNESCO Thesaurus

Serious consideration is being given to a possible revision of the "UNESCO Thesaurus: a structured list of descriptors for indexing and retrieval in the field of education, science, social science, culture and communication". The Thesaurus is a trilingual vocabulary principally used to identify and retrieve information stored in UNESBIB, the bibliographic database of the Unesco Integrated Documentation Network. It also serves to produce printed indexes to Unesco periodicals and as a reference tool for other Unesco information serves for documentation centers of its Regional Offices, affiliated non-governmental organizations and various library/information services in Member States, as each of these develop their own specializied information processing vocabularies and systems.

The Thesaurus was published in 1977 in English with French and Spanish editions in 1983 and 1984. With a view to harmonizing both the vocabulary and its structure with that of the thesauri of the international community in general and the UN system in particular, the preparation of the new edition is envisaged in two successive stages:

- Updating of the vocabulary (alphabetic display)
- Restructuring of the facets (systematic display).

(Abridged from UNISIST Newsletter 1989, No.3, p.64)

Standards for International Exchange of Bibliographic Information

A Summer School on Standards for the International Exchange of Bibliographic Information will be held August 3-19, 1990 at the School of Library Archive and Information Studies, University College London. A team of international speakers will discuss the importance of standards for the international exchange of bibliographic records. Proposed topics will include: standards for records creation (e.g. AACR, ISBD, Script creation); standards for subject access (e.g. DDC, UDC, and LCSH); standards for machine-readable records (e.g. MARC, UNIMARC, CCF, ISO 2709); and standards for the communication of machine-readable records (e.g. ISO standards for Open Systems Interconnection). The programme will involve workshops, demonstrations, discussion groups, and a programme of visits. The cost will be approximately 495 English pounds. Accomodation can be arranged at extra cost. For further details write to: Dr.I.C.McIlwaine, School of Library Archive and Information Studies, University College London. Gower Street, London WC1E 6BT, England, Tel.: 01-387-7050.N.Williamson