H. L. Somers,
University of Manchester Institute of Science and Technology

Observations on Standards and Guidelines Concerning Thesaurus Construction

Somers, H. L.: Observations on standards and guidelines concerning thesaurus construction.
In: Int. Classif. 8 (1981) No. 2, p. 69–74, 13 refs.

An attempt is made to compare the existing standards and guidelines for thesaurus construction and development, focussing particularly on the ISO, BSI standards as well as on the guidelines suggested by Aitchison and Gilchrist, and UNISIST. The different facets/aspects considered are: linguistic aspects of thesauri; formal requirements suggested by the standards/guidelines with special emphasis on problems associated with compound terms, homographs, forms of terms, etc.; semantic relationships between terms – synonymy, BT/NT, and associativity; problems peculiar to multilingual thesauri, especially the problem of inexact equivalence between terms; and presentation and arrangement of terms in a thesaurus. (acc. to Author)

1. Introduction

Preparatory work concerning a study toward "Guidelines for the Establishment of Comparison and Compatibility Matrices Between Thesauri in the Social Sciences" carried out at the University of Manchester Institute of Science and Technology (UMIST)* involved careful consideration of existing standards and guidelines. In this article a critical summary is made of the recommendations of the International Organisation for Standardization (ISO 2788) (2), and of the British Standards Institution (BS 5723) (3), AND UNESCO's own UNISIST guidelines (4).

We also took into consideration the important and influential booklet, Aitchison & Gilchrist (4). It is to be noted that the three Guidelines documents are in fact quite similar to each other, at least in spirit. BS 5723 is allegedly based on ISO 2788, though the former "has been enlarged to include procedures for dealing with compound terms". The UNISIST document differs from the others in that it deals specifically with multilingual thesauri. The Aitchison & Gilchrist booklet embodies much the same principles again, the difference being that they are treated in a more discursive, didactic manner.

In this discussion we compare each of the documents, with comments as appropriate. We give, however, little or no exemplification, particularly where this can be found in the documents themselves. For convenience the following abbreviations will be used throughout:

ISO	ISO 2788
BS	BS 5723
UNISIST	Guidelines
A & G	Aitchison & Gilchrist

Page (p.) and paragraph (para) references are to the versions cited under "References".

BS states that the two standards (i.e. BS and ISO) should not "be regarded as a set of mandatory instructions based on preferred techniques" (p. 1), wishing to avoid an "authorian approach". In Appendix D of our original report however, we recommend that, except where optional alternatives are clearly given, the recommendations set out in the standards indeed be viewed in this light (except where they are contradictory from one standard to the other) if the documents are to have any value. ISO defines as its scope the intention to "facilitate the preparation and development of thesauri . . ." (para 0), taking no account of whether the thesaurus is to be administered manually or mechanically, the subject field of the thesaurus, nor the language of the thesaurus (para 1.1.). UNISIST is to be regarded as an extension of ISO (UNISIST, para 2.1.).

2. Definitions

ISO defines *thesaurus* both in terms of its function –

> *a terminological control device used in translating from the natural language of documents, indexes, or users into a more constrained 'system language'* (para 2.1.)

and of its structure –

> *a controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge* (ibid.).

The BS definition reflects the structural aspect:

> *a means for displaying the terms in a controlled indexing language, together with indications of their a priori relationships* (para 3.4.).

Both ISO (para 2.1.) and A & G (p. 2–3) stress the assumption that a thesaurus is associated with a corpus of documents, recommending that its information content and language reflect those of the corpus, while at the same time taking into consideration potential users' information needs and language usage. A & G further continually stress that the value of a thesaurus can best be judged by its performance as an indexing or document retrieval tool.

ISO restricts its definition of a *thesaurus*:

> *a thesaurus provides words or terms to express meanings that are implied by the term relationships given in the thesaurus* (para 2.2.).

The important features are the list of terms, and the fact that they are arranged hierarchically. Thesauri in which some terms are preferred to others for indexing purposes are distinguished from those in which the terminological control is performed by some unambiguous abstract concept representation device (e.g. a concept number) (para 2.1.). ISO comments, though only in a footnote, that only the former can be maintained manually:

> *thesauri not using preferred terms require machine maintenance and retrieval* (p. 1).

BS provides further definitions:

A *document* is

> *any item (. . .) amenable to cataloguing and indexing* (para 3.1.).

* The study was undertaken for the *Division for the International Development of Social Sciences* of UNESCO in February 1981 (1). These observations on standards and guidelines formed Appendix B of that study. We gratefully acknowledge UNESCO's permission to publish this Appendix here.

An *indexing term* is

> *preferably a noun or noun phrase which represents a concept* (para 3.2.).

Indexing terms are subdivided into *perferred terms* and *non-preferred terms*, sometimes known as *descriptors* and *lead-in terms* (this is the term used by A & G) or *non-descriptors* respectively, the latter being

> *provided as users' entry points in a thesaurus or alphabetical index, the indexer or user being directed by appropriate instructions (. . .) to the preferred terms* (para 3.2.).

In this article we shall mostly use the terms *descriptor* and *non-descriptor*, though in some sections of the discussion, *preferred* and *non-preferred* term are more appropriate.

A *node label* is

> *inserted into the systematic section of a thesaurus to indicate the logical basis for subdividing a category; sometimes known as a 'facet indicator'* (para 3.3.).

Several additional definitions pertaining to multilingual thesauri are to be found in UNISIST.

A *source language*

> *serves as a starting point when a (descriptor) is translated into its nearest equivalent (. . .) in a second language* (para 3.12.),

that second language being termed the *target language*. An *exchange language* is the language used "as medium for data exchange" or used for indexing and/or retrieval in multilingual systems which use only one language for these purposes (para 3.5.). *Dominant* and *secondary* are terms relating to languages in multilingual systems which do not assign equal status to all the languages (para 3.3.).

A *coined term* is a neologism created in one language where necessary to correspond to a term recognised in another language (para 3.1.).

A *loan term* is used in a similar situation where it is not felt to be necessary or desirable to create a coined term (para 3.9.).

Feedback is

> *the art of changing the form or structure of a term in a source language in order to achieve an easier or more useful solution to a problem encountered in a target language* (para 3.6.).

The example given in UNISIST clarifies this definition: The German term LEHRERBILDUNGSGESETZ does not, when translated into English or French, provide a satisfactory indexing term, and so the original term is factored into its separate components, and its status is changed to that of a non-descriptor.

3. Linguistic aspects

In all four documents, guidelines concerning the form of terms are consistent, though BS and A & G are more thorough. ISO comments that

> *it is desirable that the descriptor should contain as few words as possible, and preferably only one* (para 3.2.1.).

Recommendations on spelling are to be found in ISO (para 3.2.3. (a)), in A & G (p. 17) and in BS (para 6.6.1.); in each case it is recommended that the most widely accepted spelling be adopted, while where there are equally acceptable variants, one be preferred as the descriptor, the other spellings being included as non-descriptors.

On the question of loanwords and translations (calques), ISO (para 3.2.3. (b)) makes no recommendation other than that cross-references between the two terms be made, while BS (para 6.6.2.) is more specific: the loanword, if it is widely accepted, should be preferred; if no accepted calque exists, the loan word (and not a coined translation) should be treated as a newly-coined English term. However, if a translation becomes more widely accepted than the foreign term, this should be adopted as the preferred term. UNISIST (para 8.3.1.) is more reticent about loaned terms and coined calques, and recommends the use of scope notes and definitions for clarification, and that in any case (para 8.3.3.) indexers, language and/or subject specialists should be consulted. Several recommendations for and examples of the creation of coined terms are given (para 8.3.3.2.).

BS (para 6.3.3.) and ISO (para 3.2.3. (c)) recommend accepted standards for guidelines on transliteration – BS 2979, BS 4280, BS 4812, ISO/R 9, ISO/R 233, ISO/R 259, ISO/R 843 (5–11) – as does A & G (p. 17). BS and ISO further note that transliterations that do not use diacritical marks are preferable. UNISIST (para 10.4.3.3.) addresses the special problem that arises where different languages have differing transliteration conventions.

The use of slang terms and jargon is commented on in BS (para 6.6.4.). Such terms may emerge as useful indexing terms, though preferably as non-descriptors associated via a 'use' relationship with a non-slang descriptor. BS also recognises that no non-slang term might exist, in which case the slang term would be an acceptable descriptor. The example given – 'hippies' – is a case in point. BS also indicates that common names should be preferred to trade names (para 6.6.5.), even, apparently, where the trade name is more commonly used. On the other hand a choice between a popular name and a scientific name (para 6.6.6.) should be decided according to the norms of the field of the thesaurus and of its users. UNISIST pays particular attention to multilingual problems concerning proper names (para 10.4.), recommending that while names of national or local institutions should be retained in their original form (though with translated forms, where they exists, as non-descriptors), names of international institutions should be expressed in the appropriate translated form. Similarly, names of historical personnages that differ from language to language should be given in 'translated' form.

Finally, ISO (para 3.2.7.), BS (para 6.3.) and A & G (p. 17) make some recommendations on the use of *abbreviations and acronyms*. The former are generally to be avoided, on the quite reasonable grounds that they may be ambiguous except in context, and that they pose filing problems. The supposed problem that

> *their recognition may be dependent on capitalisation and periods which become constraints if computer printers or other electronic data processing equipment is used* (ISO para 3.2.7.)

may well be based on an outdated and invalid view of the possibilities of computer implementation. The guidelines note that abbreviations are acceptable descriptors if they are commonly preferred by users, especially if the expanded forms would be less practical by dint of their length. Both ISO and BS state that well-established acronyms are quite acceptable as descriptors.

4. Formal requirements

On the question of *compound terms*, ISO makes few recommendations (para 3.2.1.): although

> *it is desirable that the descriptor should contain as few words as possible, and preferably only one,*

it is recognised that this practise may result in loss of clarity. Compound descriptors, when they are used, should retain natural word order. BS on the other hand has several pages of recommendations and comments on compound terms (para 7). It notes in detail some of the problems of factoring compound terms (para 7.2.), and lists several classes of term where factoring would be counterproductive, and which should therefore be retained as compounds (para 7.2.6.1.). BS also describes several situations where factoring of compound terms would be desirable (para 7.2.6.2.), though its recommendations here are rather more tentative. It is probably not desirable to list here details of these recommendations. Suffice it to say that BS addresses the problem in some depth. A & G treat compounds and factoring in equal detail.

Closely associated with questions of compound terms and factoring are the notions of pre-and post-coordination. Neither concept is particularly well explained in ISO (para 3.2.2.2.), while BS gives no explanation of the terms at all. A & G on the other hand cover the topic quite satisfactorily (p. 22–26), explaining, exemplifying, and comparing the alternatives.

The problem of *homographs* is treated only cursorily in both BS (para 6.5.) and A & G (p. 18). ISO notes the possible use of qualifiers to disambiguate homographs. Qualifiers, it is stressed, are part of the descriptor:

> *(homograph) and bracketed qualifiers form a compound descriptor* (para 3.3.1.).

Alternatively a scope note, which does not form part of the descriptor, may be used (para 3.3.2.). The term 'scope note' is defined only fleetingly in A & G (p. 19), whereas ISO (para 3.3.2.) and BS (para 6.7.) make more recommendations. A & G use the same term – 'scope note' – to cover both qualifiers and scope notes, and distinguish the two only informally (p. 19). ISO notes the further use of a scope note to indicate the history of the term's inclusion in the thesaurus, and its source (especially in the case of a neologism), while BS also suggests indications of restrictions of use, explanation of abbreviations and acronyms, and even exclusion of possible meanings. Both standards stress that scope notes should be clearly distinguished from the terms to which they are appended. ISO further recommends definitions (para 3.3.3.) and even translations (para 3.3.4) as means of disambiguating homographs. UNISIST (para 10.2.3.) mentions the interesting problem of *interlanguage homographs*, which however should not lead to confusion,

> *except in the unlikely event that the editors decide to print terms from more than one language as a single alphabetical sequence.*

Rules governing the forms of terms, especially the question of singular and plural, are given in ISO (para 3.2.4. and 3.2.5.), in BS (para 6.4) and in A & G (p. 14–15), where, in addition, a table from EJC Rules (12) is reproduced. UNISIST (para 10.1.2. and 10.1.3.) refer the reader to ISO, but note that different conventions apply for different languages. It is noted however that

> *there is no need to apply a single convention to all the languages* (para 10.1.3.).

The conventions for English are explained at length in BS (para 6.4.): in general, plurals of nouns are indicated, except in the case of non-count nouns, and the names of abstract concepts (though not those which represent classes with more than one member). The acceptability of parts of speech other than nouns as terms is discussed in BS (para 6.2.1., 6.2.2. and 6.2.3.), in ISO (para 3.2.6.) and in A & G (p. 14).

Finally, in ISO (para 3.2.8.) we find some interesting comments on problems relating to character sets. We have already mentioned ISO's preconceptions about computerisation in the comment on abbreviations (para 3.2.7.). A further statement indicates that ISO is out of date at least with respect to currently available possibilities with regard to *computer character sets:*

> *The evential use of electronic data processing equipment may entail*
> – *the use of only the upper case format for the descriptors,*
> – *avoidance of diacritical marks,*
> – *limitation of the number of characters that a descriptor may have* (para 3.2.8.).

None of the supposed entailments of computerisation of a thesaurus would be imposed in an up-to-date implementation.

Apart from these comments, ISO does make recommendations concerning the use of punctuation marks, numerals and special characters in descriptors, which are not present in the other documents.

5. Semantic relationships

In their sections on what is after all an (if not the) "indispensable function of a thesaurus", namely, "to present the interrelationships between concepts . . ." (ISO para 3.4.1.), the documents under review here converge to a much greater extent. While in the previous sections there have been only a few cases of two documents containing contradictory recommendations, it has nevertheless been the case that different aspects are given more prominence, or treated in greater depth in different documents.

In this section we summarise and compare the descriptions of the small number of widely accepted relations typically represented in thesaurus entries, as described by the documents under review. We also look especially at UNISIST, in which problems specific to multilingual thesauri of inexact term, concept and relation equivalences are discussed.

In each of the documents, the first semantic relationships to be discussed are the *equivalence relationships*, which are further subdivided into synonyms and quasi-synonyms. BS and A & G note that while in general linguistic terms true synonyms are rare, in indexing they are more common (BS para 8.2.2., A & G p. 27). This is because in indexing the meaning of any term is deliberately restricted. BS exemplifies possible causes of synonymy: terms of different linguistic origin, popular vs. scientific names, common nouns vs. trade names, variant names for new concepts, terms dating from different periods, variant spellings, terms from different cultures, abbreviations vs. full forms, and factored vs. unfactored compound terms. ISO gives similar examples under 'USE-

reference' (para 3.4.2.(a)). Some of these alternatives are dealt with in sections on the forms of terms. *Quasi-synonyms* (as distinct from synonyms) are

> *terms whose meanings are regarded as different in ordinary usage, but which are treated as synonyms for indexing purposes* (BS para 8.2.3.).

A & G (p. 28) provide examples of terms representing points on a continuum, having a significant overlap, and specific concepts subsumed (within the scope of a particular thesaurus) under broader terms. BS calls this "upward posting" (para 8.2.4.). The *quasi-synonymity relationship*, note BS and A & G, should only be used in fringe areas:

> *(It) should not be used as a means for reducing the number of (descriptors) in an indexing language* (BS para 8.2.3.).

Each of the documents notes the most commonly-used symbols for synonymity: USE, to lead from a non-descriptor to a descriptor, and its converse UF (USED FOR). UNISIST (para 4.1.) provides French and German equivalents for these and other symbols. ISO (footnote, p. 7) notes that these and other symbols will be changed when international agreement has been reached. In a multilingual thesaurus, all descriptors must be matched in each language. The problems that this entails are discussed in our original report. Different languages may have different numbers of synonyms for any given descriptor however, and while it is not necessary to establish correspondences between non-descriptors in different languages, narrower term descriptors must be matched one-to-one across the languages (cf. UNISIST para 11.2.). BS mentions a further equivalence relationship, the *Used for combination reference* (UFC). This is for

> *cases where a descriptor is used in combination with other descriptors to denote a concept represented by a non-descriptor* (para 3.4.2. (c)),

and is thus the reciprocal of a USE of semantic factoring.

The hierarchical relationship is

> *the basic relationship which most distinguishes a systematic thesaurus from other organised lists of terms* (BS para 8.3.1.).

The notions of broader term (BT) and narrower term (NT), or super- and subordinated term are used and explained in all the documents. There are two basic hierarchical relationships: the generic and the partitive. The generic relationship (for which the special symbols BTG and NTG are suggested) is that of a class or category of concepts and its members. The partitive relationship (symbols BTP and NTP) is that of a whole to its parts. Examples given in BS (para 8.3.3.) more or less relevant to this study are geographic locations, disciplines or fields of discourse, and social structures. ISO notes that if *partitive relationships* are of no significance in hierarchical retrieval,

> *it is recommended that only the generic relation be represented by hierarchical reference. In this case the part-whole-relationship is treated as associative relation* (para 3.4.3.).

BS describes a third type of relationship: the *polyhierarchical relationship*. This is where

> *a concept can logically be designated as a member of more than one category at a time* (para 8.3.4.).

This relationship is important in a multilingual thesaurus, where a given hierarchy may be regarded as 'natural' in one linguistic culture, but not in another:

> *Such fundamental differences between the categorial systems of different language users would tend to indicate that the terms in these languages either refer to different concepts, or they express the same basic concepts from such different viewpoints that the hierarchy expressed in the source language cannot be translated, as it stands, into the target language* (para 11.3.2.).

UNISIST states that in this circumstance the term should be treated as polyhierarchical.

The third, and 'vaguest' of the semantic relationships is the *associative relation*, where terms

> *are not equivalents nor do they form a hierarchy (. . .), yet they are mentally associated to such an extent that the link between them should be made explicit . . .* (BS para 8.4.1.).

The symbol RT (related term) is generally recommended. It is interesting that A & G regard the partitive relationship as associative rather than hierarchical (p. 29). A & G give some examples of other associative relationships: a property, process, attribute, application of something, property and attribute of a process, agent and instrument in a process, property of a process (p. 30). ISO gives further exemplification (para 3.4.4.): antonymity, genetic relation, cause and effect, material relation, etc. It is interesting that while ISO (para 3.4.4.) recommends that concepts related to a super-ordinated concept in the same way (i.e. terms which share a common BT) should also be related associatively, BS (para 8.4.2.) does not recommend the association of all such "sibling terms". UNISIST cautions that the validity in one language of an associative relationship in another should be determined before the translation is made. However,

> *despite this injunction, a multilingual thesaurus should usually contain a richer variety of associative relationships than a monolingual thesaurus in the same field, since it will benefit from the viewpoints of different language users* (para 11.4.2.).

6. Multilingual equivalence

UNISIST (para 9) treats in detail the most important problem for multilingual thesaurus compilers, namely the establishment of equivalent terms in different languages. The document notes that exact *equivalences* can quite often be found, which may or may not be mutually cognate, or may

> *appear to express the same concept from different viewpoints* (para 9.1.).

It might be the case however that a *general common concept* has slightly different set membership in the different languages (inexact equivalence). The solution according to UNISIST is that

> *terms which differ only in connotation should be treated, for indexing purposes, as exact equivalents* (para 9.2.).

Where terms generally refer to the same concept, but strictly one term denotes a slightly broader or narrower concept (partial equivalence), two solutions are offered (para 9.3.): the preferred solution is to regard the situa-

tion as analogous to quasi-synonymy in a monolingual thesaurus. Alternatively the terms in each language are adopted as loan terms in the other languages, and are then organised hierarchically.

Three further scenarios are covered under the heading *single-to-multiple term equivalence.*

In the first, a concept represented by a single term in one language is regarded as consisting of two or more concepts in another (para 9.4.1.). The first solution requires that the single source language term be deemed to be equivalent to a combination of the more specific target language terms. A second solution treats the single source term as a homograph, further specified with a qualifier or scope note. A third solution is a combination of the first two solutions, while the fourth, and least recommended, solution is to treat the source language term as a loan term entering into a BT-NT relationship with the native target language terms.

The second scenario is similar to the first, except that the single source language term can easily be factored semantically into terms for which exact equivalences can be established in the target language (para 9.4.2.). Three solutions, none of which is especially preferred, are offered. The first involves the factoring of the single term, with re-expression where necessary of the elements as nouns. If the original term is felt to be a likely lead-in term, it can be retained as a non-descriptor. If factoring would result in distortion, then the foreign language equivalent is expressed as a combination of the two (or more) target language terms. The third solution is to devise a coined term which will serve as an exact equivalent, adding a scope note stating that the term is a neologism, and explaining its meaning if it is not obvious.

The third type of single-to-multiple term equivalence is rather more complex. Here there exist exactly equivalent terms in both languages which enter into a hierarchical relationship. However, there also exists in one language an intermediate term for which no corresponding term in the other can be found (para 9.4.3.). Three solutions are given, of which the first is preferred. The first solution is similar to the second solution of scenario 2, namely to establish as the target language equivalent a combination of the subordinated terms, with a scope note indicating the single source language term. The second solution involves the assignment of non-descriptor status to the source language term, associating it via a USE indication with either the broader or the narrower term for which the problem of non-equivalence does not apply. In the third solution, the source language term is adopted in the target language as a loan term, as previously, accompanied by an explanatory scope note, and exact equivalence is established.

Finally, the rarely encountered situation of non-equivalence is considered (para 9.5.). This arises either where a term is so culture-bound that it is unknown to users of the target language(s), and so no translation exists, or where a scientific or technological term has only been coined in the language of the inventor, discoverer or perpetrator of the process, operation, equipment or phenomenon referred to.

The two solutions offered, neither of which is preferred, are similar to solutions suggested for parallel problems of partial equivalence, and involve again either the adoption of the term as a loan term, or the coining of a new term, in either case with an explanatory scope note.

In para 9.6., UNISIST shows how some or all of the problems mentioned can be encountered, even in the establishment of a single equivalence. A worked example is given, showing how an equivalent in French is found for the English term TEENAGERS.

7. Presentation and arrangement

While the most important aspects of a thesaurus – its linguistic and semantic content – are dealt with as described above, all the documents under review also consider the physical layout (in printed form at least) of the thesaurus to be important enough to merit several paragraphs of recommendations.

On the question of an *introduction,* ISO (para 4.1.) and BS (para 10.8.2.) are both most elaborate, and the several recommendations were incorporated in the 'checklist for Social Science Thesauri' developed as part of the original report. UNISIST has a similar rubric (para 13.2.1.), adding that for a multilingual thesaurus, the information should be given both on and in each of the languages. The introduction should also include further instructions particularly pertinent to problems of tracing translational equivalents of terms.

The main part of the thesaurus is that part in which complete information on each descriptor or concept is shown. None of the standards implies that the *systematic display* should necessarily be the main part of the thesaurus, though A & G comment that

> *early thesauri were entirely alphabetical, but (. . .) the deficiencies of this arrangement have become apparent and the systematic approach is now widely accepted* (p. 50).

ISO recommends that a thesaurus should include both a systematic and an alphabetical display (para 4.2.), while BS notes that a thesaurus which is mainly organised systematically should have a supporting alphabetical index (para 9.1.), the link between the two parts being provided by a system of address codes (para 9.3.1.). On the subject of systematic vs. alphabetical displays, BS is rather exemplificatory than advisory. It is clear in all the documents however that whatever relationships are incorporated in the thesaurus, they should all be shown in the main part (whether this be hierarchical or alphabetical), while the amount of information contained in the auxiliary parts may vary quite considerably from thesaurus to thesaurus.

Special problems of *presentation for multilingual thesauri* mentioned in UNISIST are principally those of the necessity of providing independent alphabetical indexes for each language (para 12.1.2. (a)), and those of space (on the printed page):

> *in some circumtances a need to economise on space may militate against the adoption of a form of display which would otherwise have been regarded as more acceptable on intellectual grounds, or more convenient from the user's point of view* (para 12.1.2. (b)).

UNISIST also notes the necessity, in the unlikely event of the thesaurus compilers deciding to include a multilingual alphabetical index, of disambiguating interlingual

homographs (para 10.2.3.); indeed this practise is quite strongly not recommended elsewhere (para 12.2.2.), as mentioned above. UNISIST notes that the *adoption of non-language-specific symbols* to denote the various relationships displayed in thesauri might overcome any problems relating to the fact that otherwise recommended symbols represent abbreviations selected from particular natural languages (p. 58). It is noteworthy however that some of the symbols suggested by ISO (UNISIST, p. 59) do not appear on standard typewriter keyboards. BS (p. 23) gives a slightly different set of language-independent symbols, noting however that "they have not yet been officially adopted".

Different bases for *alphabetisation* are considered. Apart from the multilingual problem of variations in 'alphabetical order' (UNISIST, para 13.2.2. (a)), A & G (p. 54) and ISO (para 4.4.) both describe two types of collating sequence, letter-by-letter and word-by-word, to which ISO adds computer sort: in the letter-by-letter sort, spaces between words are ignored, and numerals are sorted in ascending value, all other characters – except left parentheses–being ignored; in the word-by-word sort, terms beginning with a given complete word precede any terms beginning with the same sequence of letters as part of a word, and non-alphanumeric characters are treated as spaces; in the so-called computer sort, terms are sorted according to a sequence of all characters, including non-alphanumerics, determined by particular implementations, so that typically for example U.K. comes after URUGUAY (because full-stop sorts lower than 'R').

Both A & G (p. 50) and BS (para 9.2.1.) make recommendations about how information for each term should be presented, suggesting a specific order in which the different relationships should be set out. The two correspond in their recommendations: the descriptor, preceded by its address code, is followed by, in this order, scope notes, synonyms, hierarchical relations (broader before narrower), then associative relations. As A & G comment, this is in any case the most common layout.

The three standards describe in a more or less exemplificatory rather than recommendatory fashion the possibilities of including as one of the auxiliary parts of the thesaurus a graphic display (ISO para 4.3.4., BS para 9.4., UNISIST para 12.4.). Since these possibilities are so varied and in any case unsuited either to comparison or conflation, it is probably not worthwhile considering in great detail here these recommendations.

References

(1) Sager, J. C., McNaught, J.: Guidelines for the establishment of comparison and compatibility matrices between thesauri in the Social Sciences (Report for the Division for the International Development of Social Science, Unesco). CCL/UMIST Report No. 81/2. Manchester: Centre for Computational Linguistics, University of Manchester Institute of Science and Technology 1981.

(2) International Organisation for Standardization (I.S.O.): Documentation: Guidelines for the establishment and development of monolingual thesauri. ISO 2788, 1974.

(3) British Standards Institution Guidelines for the the establishment and development of multilingual thesauri. BS 5723, 1979.

(4) Austin, D., Waters, J.: UNISIST guidelines for the establishment and development of multilingual thesauri (revised text). PGI/80/WS/12. Paris: Unesco 1980.

(5) Aitchison, J., Gilchrist, A.: Thesaurus construction: a practical manual. London: Aslib 1978.

(6) British Standards Institution: Transliteration of Cyrillic and Greek characters. BS 2979, 1979.

(7) British Standards Institution: Transliteration of Arabic characters. BS 4280, 1976.

(8) British Standard Institution: The romanization of Japanese. BS 4812, 1976.

(9) International Organisation for Standardization (I.S.O.): International system for the transliteration of Slavic Cyrillic characters. ISO/R 9, 1968.

(10) International Organisation for Standardization (I.S.O.): International system for the transliteration of Arabic characters. ISO/R 233, 1961.

(11) International Organisation for Standardization (I.S.O.): Transliteration of Hebrew. ISO/R 259, 1962.

(12) International Organisation for Standardization (I.S.O.): International system for the transliteration of Greek characters into Latin characters. ISO/R 843, 1968.

(13) Engineers Joint Council (E.J.C.): Rules for preparing and updating engineering thesauri. New York: E.J.C. 1965.

Mr. H. L. Somers
UMIST. The University of Manchester
Institute of Science and Technology.
P.O.Box 88,
Manchester M60 1QD
England

WISSENSSTRUKTUREN UND ORDNUNGSMUSTER

(Structures of Knowledge and Patterns of Order)

Proceedings der 4. Fachtagung der Gesellschaft für Klassifikation e.V., Salzburg, 16.-19.April 1980. Red.: Wolfgang Dahlberg. Frankfurt/Main INDEKS VERLAG 1980. 368 p., DM 65.- (ca 27.- $)

Contains the six plenary lectures (by Walter Oberschelp, Carsten Bresch, Rudolf Haase, Wilhelm Totok, Janos S. Petöfi and Ingetraut Dahlberg) with extensive discussions as well as 15 papers of session meetings (by H.G. Körner, W.Zwirner, P.O.Degens, F.Jochum, M.Koch, S.Rösch, F.Seitelberger, Ch.Weinberger, Arno Müller, O.Sechser, J.Hölzl, H.Mönke, J.Panyr, Ota Weinberger, B.Maassen, also with discussions) on structures of knowledge and patterns of order as these may be found in six main areas of human knowledge and activity. Concludingly reports on the SIGs.