THESAURUS
A tool of Information Organization

Introduction

- The efficiency of information retrieval system largely depend on the indexing language adopted, while that of the latter depends on its capability to handle two fundamentally different but interdependent types of relationship between the terms used for representing the subject matter of the document.
- Though, thesaurus has been conceived mainly in the context of post-coordinate indexing system, it can be used for pre-coordinate indexing systems as well.

Background

- The word Thesaurus, which is of Greek origin, literally means “ *treasury or storehouse of Knowledge*”. But in modern usage, it denotes a list of terms arranged according to their relationship of ideas.
- It was Peter Mark Roget who first conceived the idea of such a compilation and brought out in 1852 his Thesaurus of English words and Phrases for the benefit of writers for looking for appropriate words to express their ideas.

Background (contd.)

- Helen Brownson is said to be the first person who used the term ‘ *Thesaurus*’ in the context of information retrieval in a paper presented in 1957 at the Dorking Conference in Classification Research, and Hans P. Luhn was possibly the first person to think in terms of ‘*Information Retrieval Thesaurus*’.
- The first Thesaurus used in an information retrieval system was developed by Du Pont in USA around 1959 and since then no. of thesauri have been brought out.

Definition

"A *thesaurus* may be defined either in terms of its function or its structure. In terms of function, a thesaurus is a terminological control device used in translating from the natural language of documents, indexers or users into a more constrained "system language" (documentation language, information language). In terms of structure, a thesaurus is a controlled and dynamic vocabulary of a controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge."

Definition (contd.)

- “A compilation of words and phrases showing synonymous, hierarchical, and other relationships and dependencies, the function of which is to provide a standardized vocabulary for information storage and retrieval.”
- “A controlled vocabulary arranges in a known order in which equivalence, homographic, hierarchical, and associative relationships among terms are clearly displayed and identified by standardized relationship indicator, which must be employed reciprocally.”

Definition (contd.)

A thesaurus in the field of information storage and retrieval is a list of terms and/or of other signs (or symbols) indicating relationships among these elements, provided that the following criteria hold:

(a) the list contains a significant proportion of non-preferred terms and/or of preferred terms not used as descriptors;

(b) terminological control is intended.

Brief History

- 1959 – the Engineering Information Center of E. I. Dupont de Nemours developed the first true thesaurus
- 1960 – the Armed Services Technical Information Agency (ASTIA) produced the *Thesaurus of ASTIA Descriptors*
- 1961 – the American Institute of Chemical Engineers (AIChE) published the *Chemical Engineering Thesaurus*
- 1964 – the Engineers Joint Council (EJC) published the *Thesaurus of Engineering Terms*
- 1967 – *Thesaurus of Engineering and Scientific Terms* (TEST)

Brief History (contd.)

- 1967 – the Committee on Scientific and Technical Information (COSATI) published the first set of guidelines for thesaurus construction
- 1970 – Unesco *Guidelines for the Establishment and Development f Monolingual Scientific and Technical Thesaurus*
- 1974 – ANSI (American National Standards Institute) Z39.19 [a US national standard for thesaurus construction]
- 1974 – the first international standard for thesaurus construction – ISO 2788

Purpose and Use of Thesaurus

- "Its purposes are to promoted consistency in the indexing of documents, predominantly for post-coordinated information storage and retrieval systems, and to facilitate searching by linking entry terms with descriptors"
- Four principal purposes are served by a thesaurus:

  a) *Translation*. To provide a means for translating the natural language of authors, indexers, and users into a controlled vocabulary used for indexing and retrieval.

Purpose and Use of Thesaurus (contd.)

b) Consistency: To promote consistency in the assignment of index terms.

c) *Indication of Relationships*: To indicate semantic relationships among terms.

d) Retrieval: To serve as a searching aid in retrieval of documents.

Difference between List of Subject Heading and Thesaurus

- *List of subject heading* (LSH) is a complete list of names of subjects usually arranged in alphabetical order whereas Thesaurus is a list of terms arranged in a helpful order, in other words, it is a compilation of all isolate ideas that occur within a subject or group of subjects represented by appropriate isolate terms arranged in alphabetical order.

- *LSH* were designed in view the needs of pre-coordinate indexes whereas *Thesaurus* was designed to meet the special needs of the post-coordinate indexes.
- In *LSH*, the emphasis is on references from broader to narrower subject or downward references whereas in *Thesaurus*, thesauri have more elaborate network of referencing, the direction of each type of reference is clearly indicated.

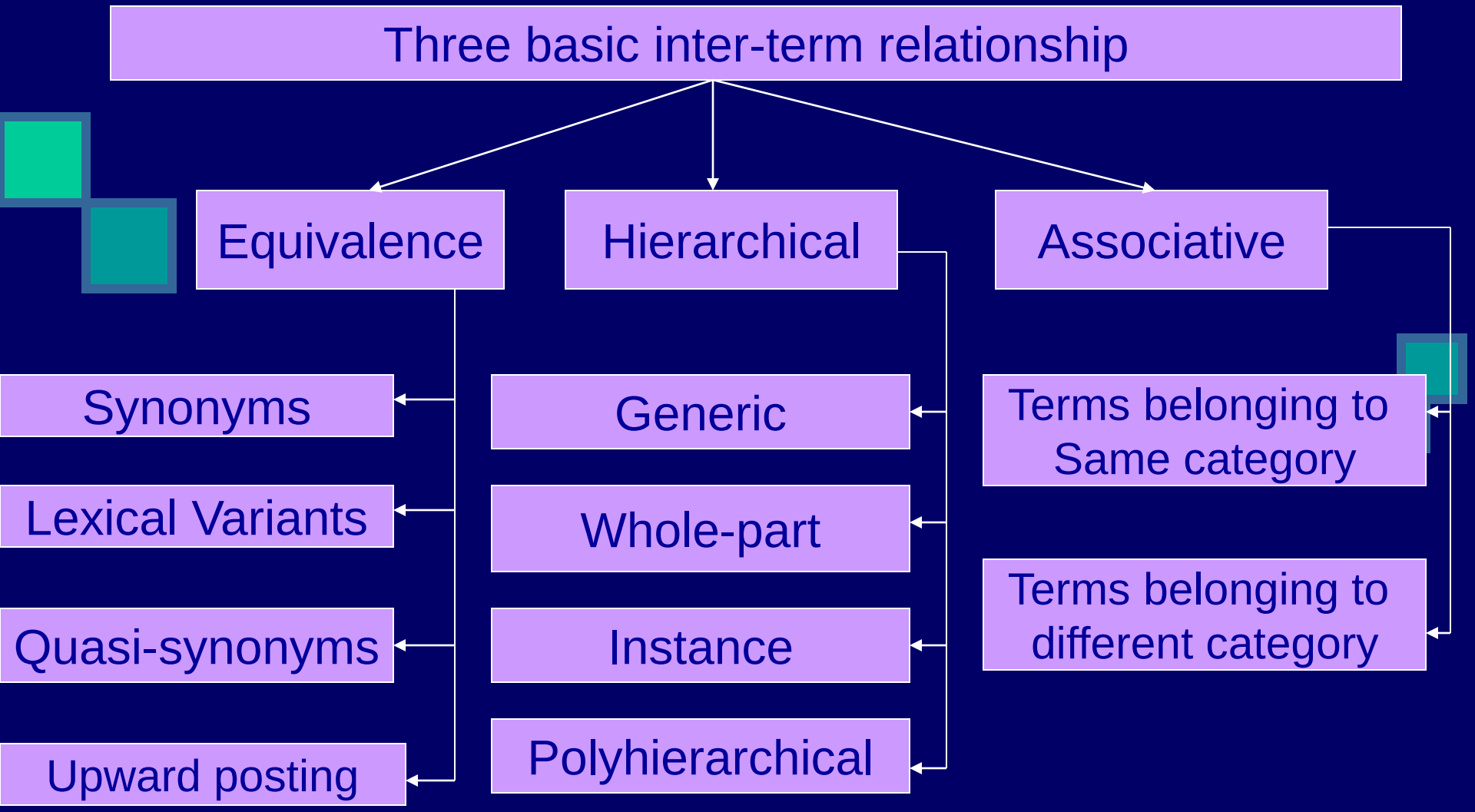
Structure and Relationship

There are two broad types of relationships within a thesaurus

the semantic links between individual terms

how the terms and their inter-relationships to the overall Structure of the subject field

Basic Thesauras Relationship



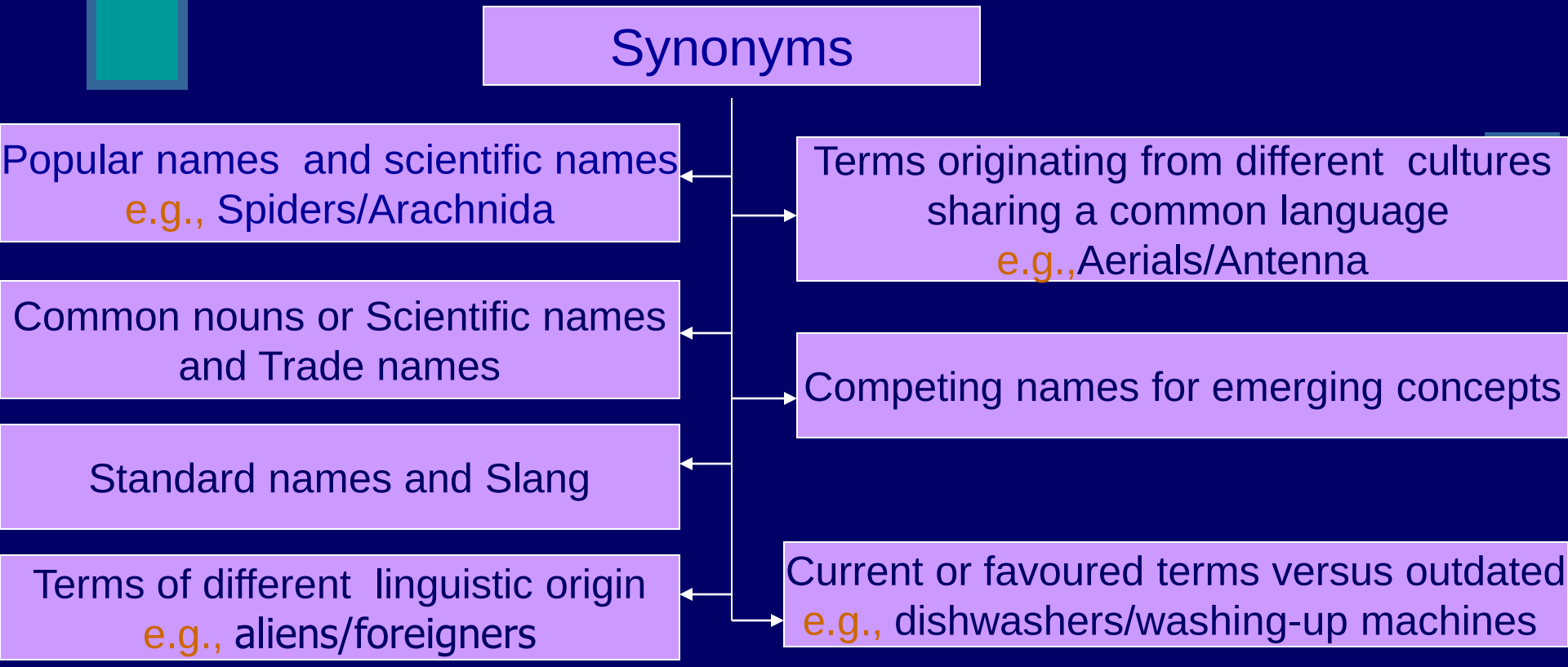
A) Equivalence Relationship

- The relationship between preferred and non-preferred terms where two or more terms are regarded, for indexing purposes, as referring to the same concept.
- The equivalence relationship includes synonyms, lexical variants, quasi-synonyms and upward posting.

1. Synonyms

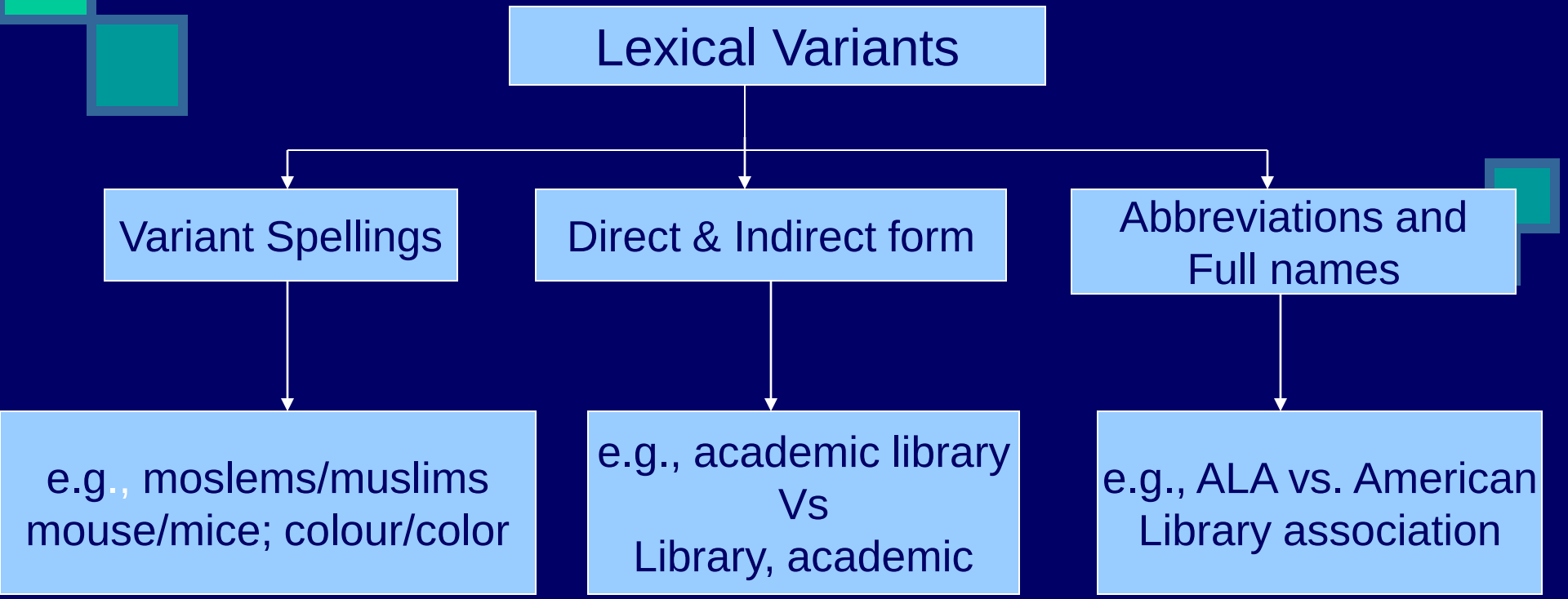
(terms whose meanings can be regarded as the same in a wide range of contexts, so that they are virtually interchangeable.)

There are several types of Synonyms:



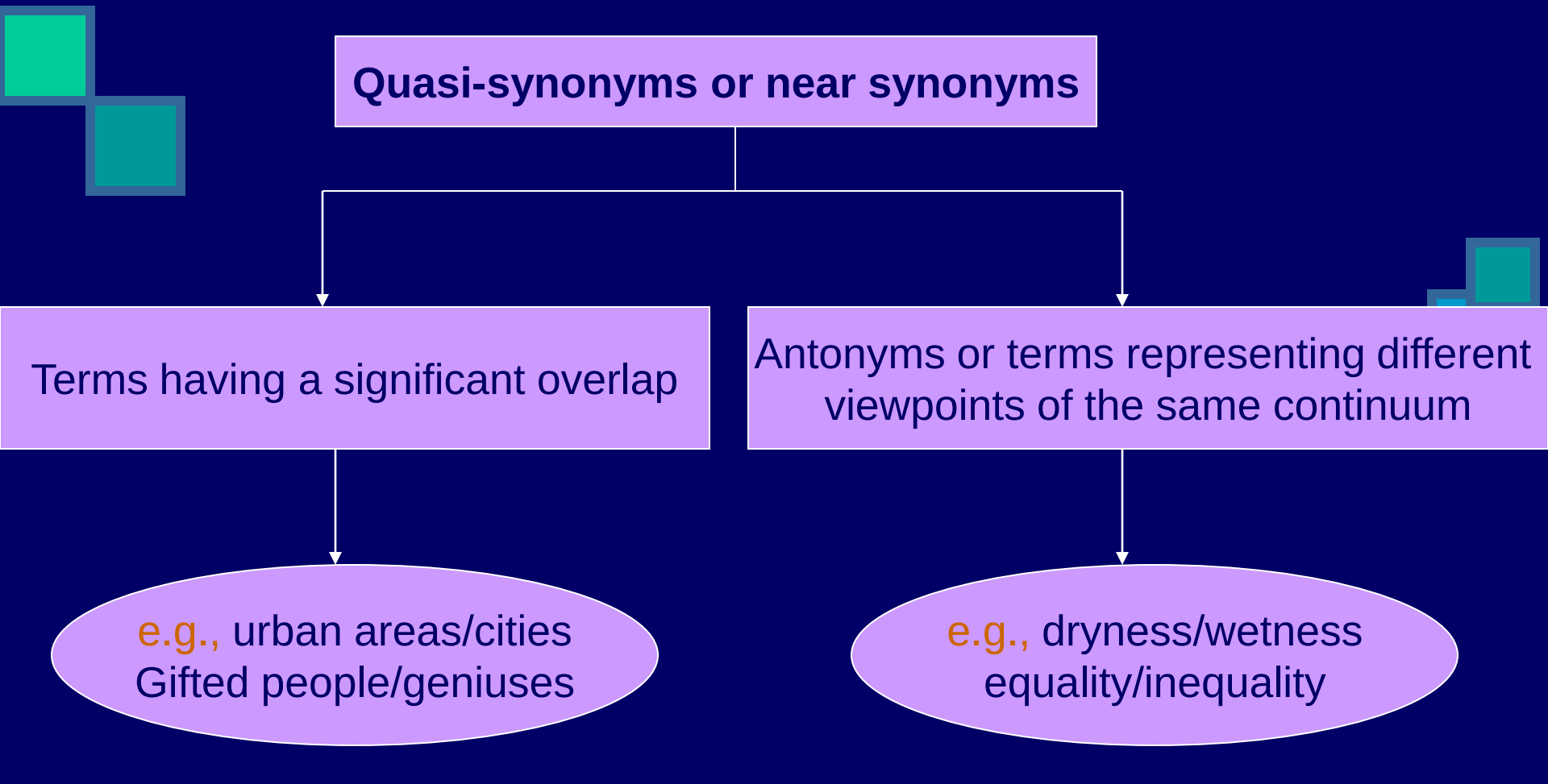
2. Lexical Variants

(different word forms for the same expressing,such as spelling, grammatical variation. Irregular plurals, Direct versus indirect order, and abbreviated formats)



3. Quasi-synonyms or near synonyms

(terms whose meanings are generally regarded as different in ordinary usage, but they are treated as though they are synonyms for indexing)



4. Upward posting

(is a technique which treats narrower terms as if they are equivalent to, rather than a species of, their broader terms. The effect is to reduce the size of the vocabulary.)

For example:

1. SOCIAL CLASS

   UF Elite

   Middle class

   Working class....

   2. Elite

      USE SOCIAL CLASS

B) Hierarchical Relationships

- This relationship shows levels of superordination and subordination. The superordinate term represents a class or whole, and the subordinate terms refer to its members or parts.
- This relationship is used to locating broader and narrower concepts in a logically progressive sequence.
- The relationship is reciprocal and is set out in a thesaurus using the following conventions:

  BT (Broader Term)

  NT (Narrower Term)

Hierarchical Relationships (contd.)

For example:

Public Libraries

BT Libraries

Libraries

NT Academic Libraries

Children's Libraries

Public Libraries ……

- The hierarchical relationship include the generic relationship, the hierarchical whole-part relationship, the instance relationship and the polyhierarchial relationship.

1. Generic/species relationship

Which identifies the link between a class or category and its members or species. The relation is also known as the inclusion relationship.

For example: VERTEBRATA

NT Amphiblia
Mammalia
Aves
Pisces
Reptilia

2. Whole/part relationship

Systems and organs of the body

Geographical location
(e.g., Taipei / Ta-an District)

Discipline or field of study
(e.g., Chemistry / Organic chemistry)

Hierarchical social structure
(e.g., army and its rank system)

3. Instance relationship

- This is the relationship between a general category of things and events, expressed by a common noun, and an individual instance of that category, the instance then forming a class of one which is represented by a proper name.
- For example: SEAS

  NT Baltic Sea
  Caspian Sea
  Mediterranean Sea

4. Poly-hierarchical relationships

- This the relationship between the term and its two or more superordinate terms is said to be polyhierarchical relationship.
- Some terms may belong to more than one hierarchy and consequently, may be related to more than one broader term and more than one set of narrower terms. Such relationship is called poky-hierarchical relationship.
- For example:

C) Associative Relationship

- The relationship is found between terms which are closely related conceptually but not hierarchically and are not members of an equivalence set.
- The relation is reciprocal, and is distinguished by the abbreviation "RT" (Related Terms)

e.g.,

TEACING

RT Teaching aids

TEACHING AIDS

RT Teaching

Associative Relationships (contd.)

Two types of associative relationship:

- Terms belonging to the same category (e.g., motorcycle / bicycle)
- Terms belonging to different categories
  - Whole-part (e.g., buildings / doors)
  - A discipline and the objects studied (e.g., ethnography / primitive societies)
  - An operation or process and the agent or instrument (e.g., motor racing / racing cars)
  - An occupation and the person in that occupation (e.g., accountancy / accountants)
  - An action and the product of the action (e.g., publishing / music scores)

Associative Relationships (contd.)

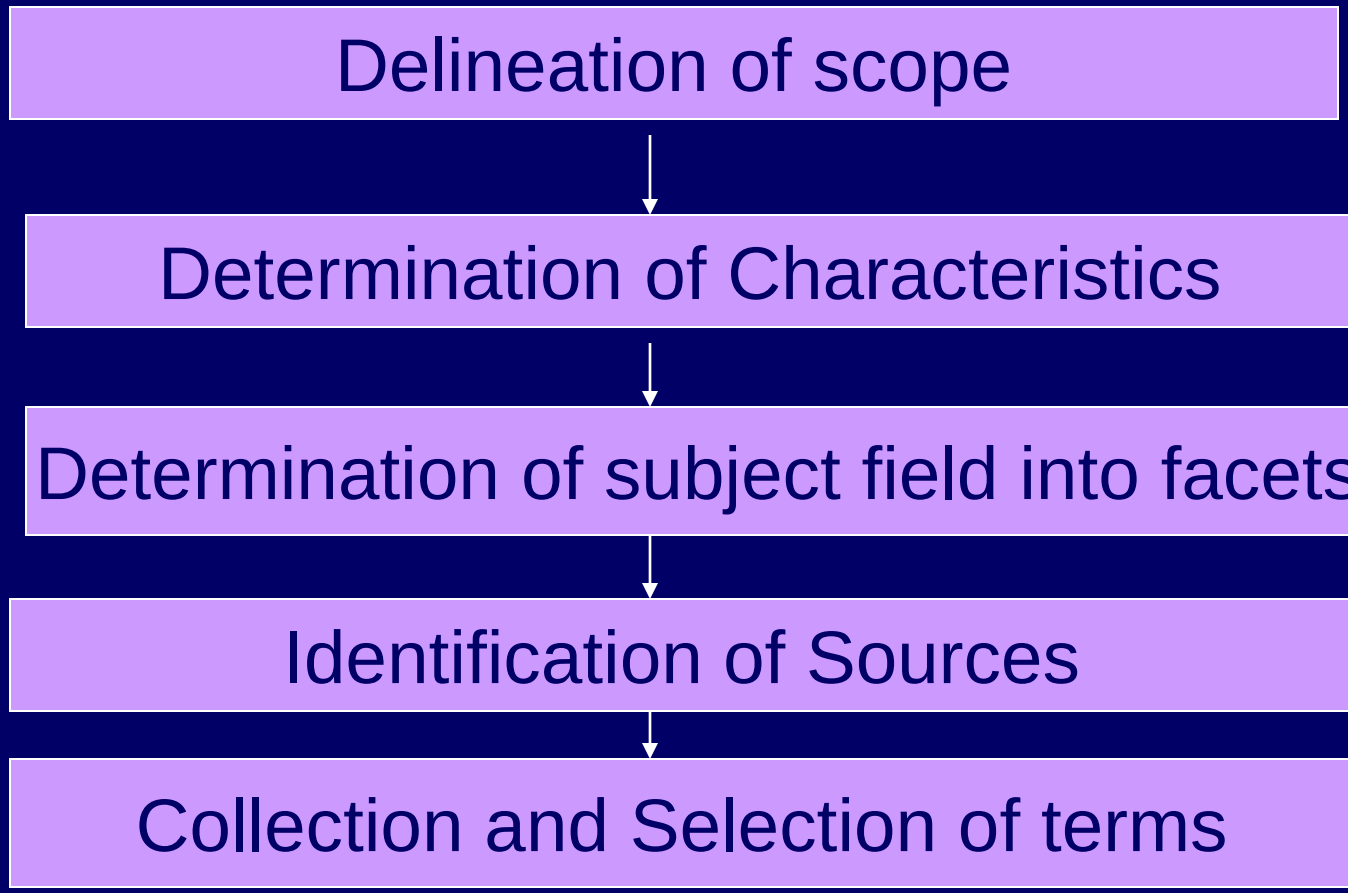
- Terms belonging to different categories (contd.)
  - An action and its patient (e.g., data analysis / data)
  - Concepts related to their properties (e.g., women / femininity)
  - Concepts linked by causal dependence (e.g., injury / accidents)
  - A thing or action and its counter-agent (e.g., pests / pesticides)
  - A raw material and its product
  - An action and a property associated with it (e.g., precision measurement / accuracy)
  - A concept and its opposite (e.g., single people / married people)

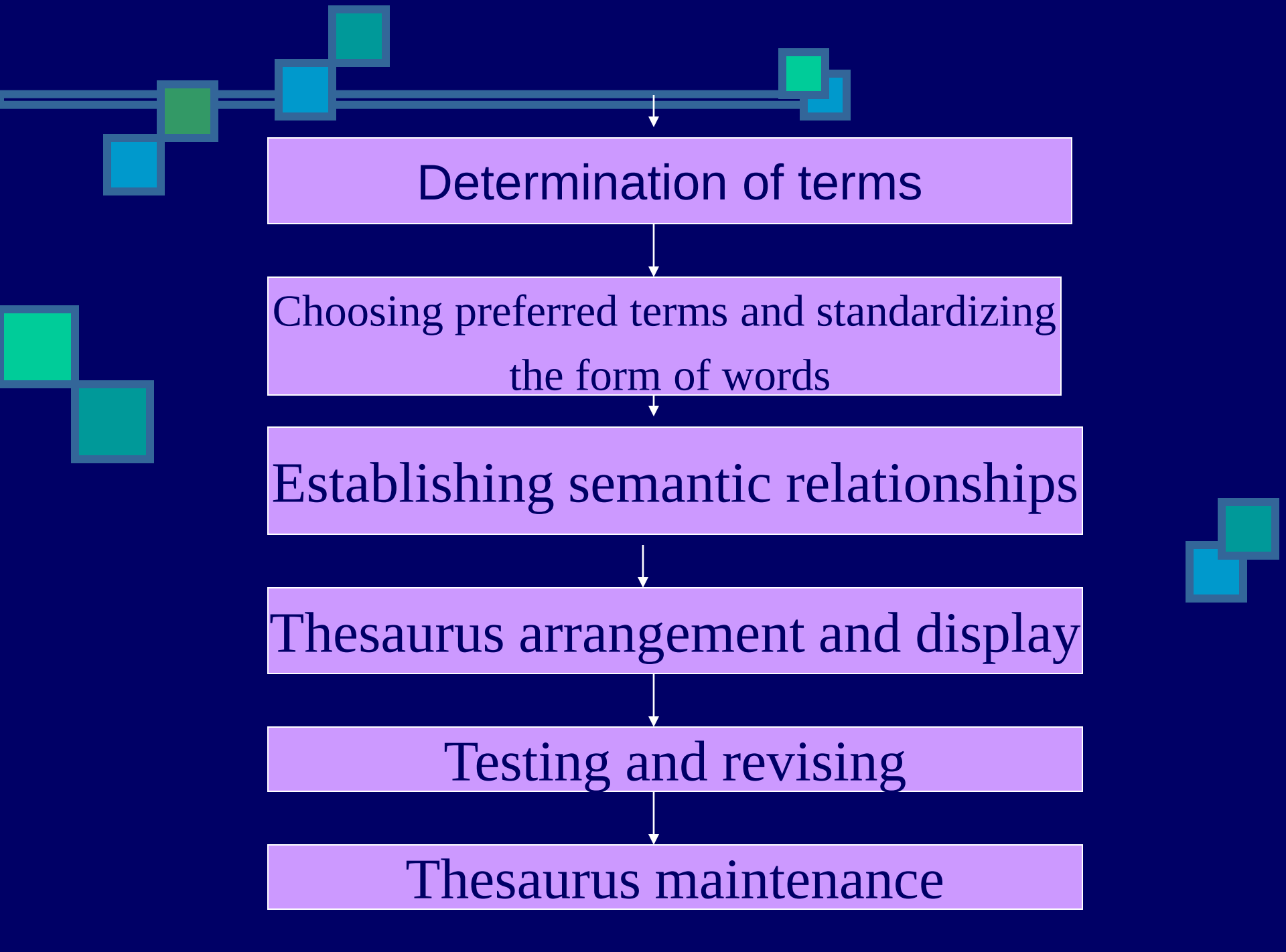
Format of IR Thesaurus

- An IR thesaurus may be arranged and presented in one or more of the following methods:

1. Alphabetical - in which descriptors and cross references are arranged in alphabetical order.
2. Systematic or Classified - in which descriptors are arranged in their hierarchical order with level of hierarchy represented by indentions, dashes, dots, etc.
3. Graphic – in which the hierarchy is shown by a tree or an arrowgraph

Essential Steps in the construction of Thesaurus




Determination of terms
Choosing preferred terms and standardizing the form of words
Establishing semantic relationships
Thesaurus arrangement and display
Testing and revising
Thesaurus maintenance

Advantages of Thesaurus

- It effect vocabulary control in the language being used in information retrieval;
- It helps an indexer in selecting preferred terms;
- It provides more access points;
- It enables the searcher to find out not only information on a specific topic, but also, on all related topics;
- By using indexing terms and search terms from the same thesaurus, the speed of retrieval can be increased;
- It helps in obtaining high recall ratio and high precision ratio in information search

Examples

- Unesco Thesaurus: A Structured List of Descriptors for Indexing and Retrieving Literature in the Fields of Education, Science, Social and Human Science, Culture, Communication and Information.
- Thesaurus of ERIC Descriptors
- Thesaurus of Sociological Research Terminology
- Thesaurus of Sociological Indexing Terms

Examples (contd.)

- Social Science and Business Microthesaurus: A Hierarchical List of Indexing Terms Used by NTIS
- Political Science Thesaurus
- SPINES Thesaurus: A Controlled and Structured Vocabulary of Science and Technology for Policy Making
- Thesaurus of Psychological Index Terms

Examples (cont.)

- Thesaurus of Engineering and Scientific Terms (TEST)
- INSPEC Thesaurus
- NASA Thesaurus
- Thesaurus of Computing Terms
- Thesaurus of Scientific, Technical and Engineering Terms
- International Road Research Documentation (IRRD) Thesaurus

Examples (cont.)

- ASIS Thesaurus of Information Science and Librarianship
- Thesaurus of Information Science Terminology
- Food: Multilingual Thesaurus
- Thesaurus of Agricultural Terms
- The ISDD Thesaurus. Keywords Relating to Non-Medical Use of Drugs and Drug Dependence

References:

- **Aitchison (Jean), Gilchrist (Alan) and Bawden (David).** *Thesaurus construction and Use; a practicle manual.* 4th edn. Chicago: Fitzroy Dearborn Publisher, 2000.
- **Ghosh (S. B) and Satpathi (J. N) ed.** *Subject Indexing systems: concepts, methods and techniques.* Calcutta: IASLIC, 1998.

Glossary

- Preferred term: Preferred terms is the one chosen to represent the concept in indexing
- Non-preferred terms : Non-preferred terms is the one not selected.
- Descriptor: an elementary term, notion or other string of symbol used to identify a subject.
- Subordinate term: Subordinate term refer to its members or parts
- Super ordinate term: Super ordinate term represent a class or whole which is high-hierarchical in order.