3 Thesaurus

Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Dealing with word meanings in information retrieval

- Problem: The same meaning can be expressed using different terms
  - synonyms
  - homonyms
  - related terms
- How can it be achieved that for the same meaning the identical terms are used in the index and the query?

Thesaurus

- A thesaurus is a sorted composition of terms and their descriptors that can be used for indexing, storing and retrieval of information in a field of documentation.
- A thesaurus contains
  - terms
  - relationships between terms

Thesaurus - Definition

- Ein Thesaurus [...] ist eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient
- Er ist durch folgende Merkmale gekennzeichnet:
  - Begriffe und Bezeichnungen werden eindeutig aufeinander bezogen (terminologische Kontrolle) indem
    - Synonyme möglichst vollständig erfasst werden
    - Homonyme und Polyseme besonders gekennzeichnet werden,
    - für jeden Begriff eine Bezeichung (Vorzugsbenennung, Begriffsnummer oder Notation) festgelegt wird, die den Begriff eindeutig vertritt,
  - Beziehungen zwischen Begriffen (repräsentiert durch ihre Bezeichnungen) werden dargestellt.

Quelle: DIN 1463 – Erstellung und Weiterentwicklung von Thesauri

Types of Thesauri

Two kinds of thesauri can be distinguished

- **Thesauri with preferred terms**
  - From the terms with the same or nearly the same meaning only one is allowed for indexing. Preferred terms are also called descriptors.
- **Thesauri without preferred terms**
  - Terms with similar meaning are collected in equivalence classes (sometimes called synonym sets or synsets). All terms can be used for indexing.

preferred term = Vorzugsbezeichnung

Thesauri in the Web

Web Thesaurus Compendium:

http://www.ipsi.fraunhofer.de/~lutes/thesoecd.html

Examples:

Thesauri with preferred terms

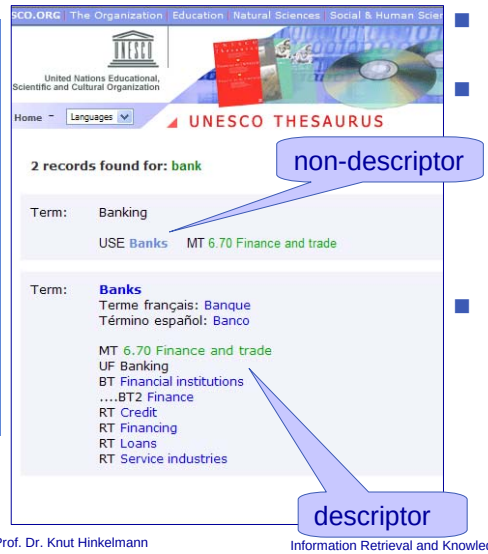
- UNESCO Thesaurus
  http://www.ulcc.ac.uk/unesco/
- Standard Thesaurus Wirtschaft
  http://www.gbi.de/thesaurus/

Thesauri without preferred terms

- Wordnet (A lexical datebase for the English language)
  http://wordnet.princeton.edu/
- Open Thesaurus
  http://www.openthesaurus.de

3.1 Thesaurus with preferred terms

Example: Unesco Thesaurus



- Terms are represented as descriptors and non-descriptors
- Descriptor
  - A descriptor, also called preferred term, is the term to be used to represent a concept when indexing documents and formulating queries
  - A descriptor contains relationships to other descriptors/terms
- Non-descriptor
  - A non-descriptor, also called forbidden term, is a term designating a concept very close to that represented by a descriptor.
  - It contains a reference to the corresponding descriptor as the only relationship

Relationships between terms

- Descriptors contain relationships to other descriptors
  - **Hierarchical** relationships, which link terms to other terms expressing more general and more specific concepts - i.e. broader terms (BT) and narrower terms (NT).
  - **Associative** relationships, which link terms to similar terms (related terms) where the relationship between the terms is non-hierarchical. Related terms are indicated by the prefix RT.
  - **Equivalence** relationships, which link "non-preferred" terms to synonyms or quasi-synonyms which act as "preferred" terms. Non-preferred terms are indicated by the prefix UF.
- A descriptor can contain additional information
  - Explanations of the intended use of the descriptor
  - Group (Microthesaurus) the descriptor belongs to
  - Lingustic equivalence, which designates the same concept in different languages for multilingual thesauri

Relations: German and English

German		English	
Abbr.	**Denomination**	**Abbr.**	**Denomination**
Hierarchy Relations			
TT	Top Term (allgemeinster Begriff)	TT	Top term
OB	übergeordneter Begriff (Oberbegriff)	BT	Broader term
UB	untergeordneter Begriff (Unterbegriff)	NT	Narrower term
Hierarchy Relations distinguishing between Abstradtion and Aggregation			
OA	Oberbegriff Abstraktionsrelation	BTG	Broader term generic
UA	Unterbegriff Abstraktionsrelation	NTG	Narrower term generic
SP	Verbandbegriff	BTP	Broader term partitive
TP	Teilbegriff	NTP	Narrower term partitive
Equivalence Relations and Associations			
BS	Benutztes Synonym oder Quasi-Synonym	USE	Use
BF	Benutzt für Synonym oder Quasi-Synonym	UF	Used for
VB	verwandter Begriff	RT	Related term
BK	Benutzte Kombination von Einfachdeskriptoren	USE	Use
KB	Benutzt in Kombination von Einfachdeskriptoren	UFC	Used for combination

Quelle: DIN 1463 – Erstellung und Weiterentwicklung von Thesauri

Equivalence Relation - Synonyms

- Semantic Equivalence is a relation between terms with (nearly) the same meaning. It is expressed by two symbols:
- **USE** – is used in non-descriptors and related to the corresponding descriptor
  - Example

    Cars
    USE Motor vehicles
- **UF** (= Used For) is used in descriptors and refers to synonymous non-descriptors
  - Example

    Motor vehicles
    UF Cars

Descriptors and Non-Descriptors

- Descriptors
  - may have zero, one or more non-descriptors corresponding to it
  - have relations to other descriptors
- Non-descriptor
  - must refer to one descriptor only (relation USE)
  - do not have any other relation
- Example from the UNESCO thesaurus:

Descriptor:

Motor vehicles
MT 6.60 Equipment and facilities
UF Automobiles
UF Cars
UF Trucks
BT Vehicles
RT Road Engineering
RT Road Transport

Non-Descriptors:

Automobiles
USE Motor vehicles

Cars
USE Motor vehicles

Trucks
USE Motor vehicles

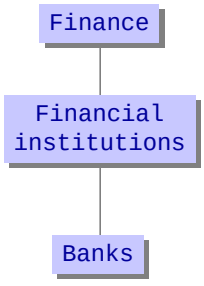
Hierarchy

- In general, a hierarchy is represented by two relations
- **BT** (= Broader Term) relates a descriptor to a more generic descriptor
  - Example:

Banks
BT Finanical institutions
BT2 Finance

- **NT** (= Narrower Term) relates a descriptor to a more specific descriptor
  - Example:

Financial institutions
NT Banks
BT Finance



- In the UNESCO thesaurus, a digit to the right of the symbols BT or NT indicates the number of hierachical levels separating the descriptors

Specific Hierachies

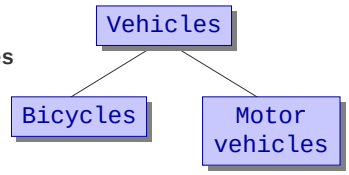
There are thesauri that distinguish between different types of hiearchies

- specific vs. generic terms: The narrower term is more specific than the broader term
  - Example:

Vehicles
NTG Motor vehicles
NTG Bicycles

Motor vehicles
BTG Vehicles

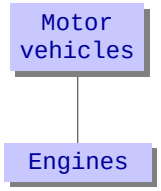
Bicycles
BTG vehicles



- partitive relation: the narrower terms is part of the broader term
  - Example:

Motor Vehicles
NTP Engines

Engines
BTP Motor Vehicles



Association RT

- RT (= Related Term) is a relation between two descriptors that is neither hiearchical nor an equivalence relation.
- There are different kinds of relations that can be expressed as association relation, e.g.
  - Descriptors that are at the same level in a hierarchy
    `Diesel engine` **RT** `Otto engine`
    `Apple` **RT** `Pear`
  - Descriptors that are part of a common thing
    `Solothurn` **RT** `Aargau`
  - Antonym (opposite)
    `Heat` **RT** `Cold`
  - Successor relation
    `Father` **RT** `Son`
  - functional or causal relation
    `Book` **RT** `Reading`

Structure of the Thesaurus

Example: subject field and microthesauri

5. Information and communication
- **5.05 Information sciences**
- **5.10 Communication research and policy**
- **5.15 Information management**
- **5.20 Information industry**
- **5.25 Documentary information systems**
- **5.30 Information sources**
- **5.35 Documentary information processing**
- **5.40 Information technology (software)**
- **5.45 Information technology (hardware)**

- The UNESCO thesaurus is organised into subject fields and microthesauri
- Field names
  - A field is a grouping of microthesaus
  - A field name is preceded by a one-digit serial number
- Microthesaurus names
  - A microthesaurus is a grouping of descriptors and non-descriptors
  - A microthesaurus name is preceded by a three-digit serial number, the first digit is the number of the subject field to which the microthesaurus belongs

Other Descriptor Information

Term: **Knowledge** [315]
Terme français: Connaissance
Término español: Conocimiento
Русский термин : Знания

SN *Information that is presented within a particular context, yielding insight on application in that context, by members of a community.*
MT 3.15 Philosophy and ethics
BT Epistemology [305]
NT Sociology of knowledge [79]
NT Structure of knowledge [33]
NT Traditional knowledge [295]
RT Information [1]
RT Know-how transfer [48]
RT Science of science [2]

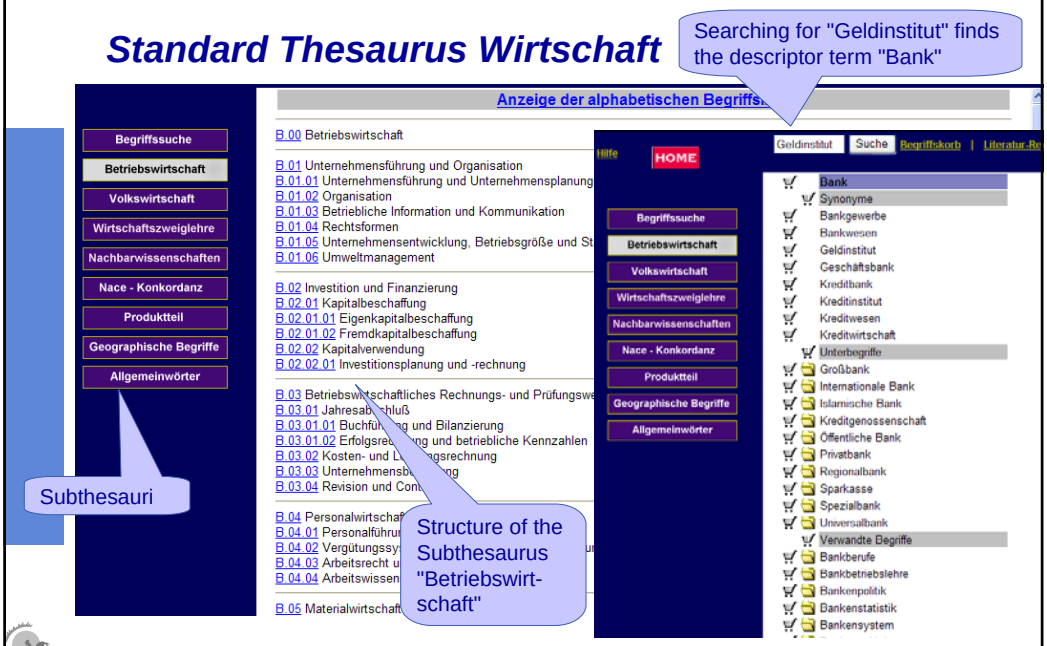
Term: **Information** [1]
Terme français: Information
Término español: Información
Русский термин : Информация

SN *Data that has been organized in such a way that it achieves meaning, in a generalized way.*
MT 5.05 Information sciences
NT Communication information [13]
NT Cultural information [110]
NT Educational information [585]
NT Scientific information [760]
....NT2 Energy information [28]
....NT2 Environmental information [92]
....NT2 Science popularization [462]
NT Social science information [233]
....NT2 Economic information [72]
....NT2 Political information [2]
....NT2 Social information [5]
RT Information transfer [315]
RT Information users [134]
RT Knowledge [315]

Descriptors in the UNESCO thesaurus also contain:

- Explanation
  - Explains the use for which a descriptor is intended
  - explanations in the UNESCO thesaurus are called Scope Notes **SN**
- Inclusion
  - Reference between a descriptor and the microthesaurus to which it belongs
  - shown by the symbol **MT**
- Linguistig equivalence
  - Relation between descriptors designatingt he same concept in different languages
  - Shown by the symbol of the language indicators

Standard Thesaurus Wirtschaft

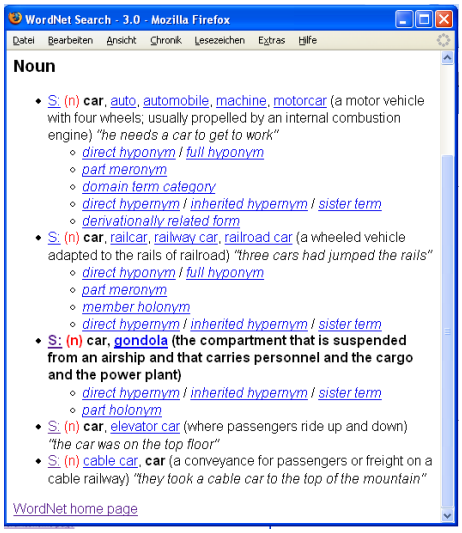


3.2 Thesauri without preferred terms

- Terms with similar meaning are represented as equivalence classes.
- Example: WordNet
  - In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets),
  - each synset expresses a distinct concept
  - Synsets are interlinked by means of conceptual-semantic and lexical relations.

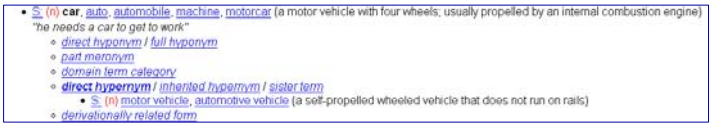
http://wordnet.princeton.edu/

Relations in WordNet

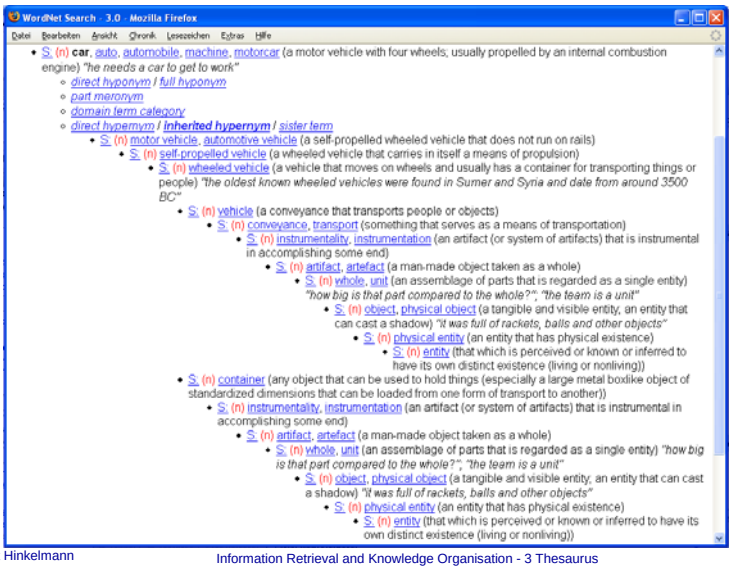


- For each synset there are a number of relations to other synsets, e.g.
  - hyponym: more specific concepts (corresponds to narrower term NT)
  - hypernym: more general concepts (opposite of hyponym; corresponds to broader term BT)
  - part meronym: consituent parts of the concept (corresponds to narrower term partitive NTP)
  - holonym: opposite of meronym
  - domain category: classes the concept belongs to

WordNet: Displaying the value of relations



WordNet: Displaying the value of relations



Example: OpenThesaurus

OpenThesaurus is an open source thesaurus for the German language

3.3 Possible uses of a Thesaurus

Index with Controlled Vocabulary

- **Use thesaurus for indexing**
  - Providing a *controlled vocabulary* for *manual indexing*
  - storing only preferred terms (descriptors) in the index, e.g. in attribute „keyword“
- **Use Thesaurus for retrieval**
  - User can use thesaurus to *formulate a query:*
    - find preferred terms
    - find broader or narrower terms if query is not successful

Fulltext search

- **Use thesaurus for indexing**
  - automatically store all synonyms as index terms
  - Thesaurus may still be helpful at retrieval e.g. to find broader terms, narrower terms, related terms
- **Use theaurus for retrieval**
  - Index contains only term occuring in the documents
  - User can use thesaurus to *refine a query:* find synonyms, broader terms, narrower termsor related terms if query is not successful

Use of a Thesaurus (2)

- The Thesaurus can be used by humans or automatically
  - Human
    - use thesaurus as a reference book
    - electronically or conventionally (book)
  - Retrieval system
    - The system can suggest synonyms, broader terms or narrower terms automatically
  - Indexing system
    - automatically find synonyms and preferred terms

Example fromthe UNESCO Thesaurus

Motor vehicles [39]
Terme français: Véhicule à moteur
Término español: Vehículo automotor
Русский термин : Автомашины

MT 6.60 Equipment and facilities
UF Automobiles
UF Cars
UF Trucks
BT Vehicles [5]
RT Road engineering [25]
RT Road transport [44]

- The figures shows a descriptor in the UNESCO thesaurus
- To the term „Motor vehicles“ there various synonyms, broader terms and related terms
- Use for indexing:
  - The index must not contain the nonb-preferred terms „Automobiles“, „Cars“, „Trucks“ but only „Motor vehicles“
- Use for keyword search:
  - Searching for „Cars" does not provide a result.
  - Looking up the thesaurus, we find that „Motor vehicles“ is the corresponding descriptor term which is used as index term.
- Use for fulltext search:
  - If searching for „Motor vehicles" provides too many results, we can use the thesaurus to find alternative search terms.

Example from the Standardthesaurus Wirtschaft

- Bank
  - Synonyme
  - Bankgewerbe
  - Bankwesen
  - Geldinstitut
  - Geschäftsbank
  - Kreditbank
  - Kreditinstitut
  - Kreditwesen
  - Kreditwirtschaft
  - Unterbegriffe
  - Großbank
  - Internationale Bank
  - Islamische Bank
  - Kreditgenossenschaft
  - Öffentliche Bank
  - Privatbank
  - Regionalbank
  - Sparkasse
  - Spezialbank
  - Universalbank
  - Verwandte Begriffe
  - Bankberufe
  - Bankbetriebslehre
  - Bankenpolitik
  - Bankenstatistik
  - Bankensystem
  - Bankgeschichte
  - Bankrecht
  - Finanzintermediär
  - Zentralbank

- The figures shows a descriptor in the thesaurus „Wirtschaft“
- To the term „Bank“ there various synonyms, narrower terms and related terms
- Use for indexing:
  - The index only contains the descriptor term „Bank“, but not the corresponding non-descriptors
- Use for keyword search:
  - Searching for "Kreditinstitut" does not provide a result.
  - Looking up the thesaurus, we find that „Bank“ is the corresponding descriptor term which is used as keyword.
- Use for fulltext search:
  - If searching for "Bank" provide too many results, we can use the thesaurus to find alternative search terms.

Maintenance of a Thesaurus

- Building and maintaining a thesaurus is requires expertise and is time-consuming
  - What terms are descriptors?
  - Are all synonyms included?
  - What is the correct relation between terms?
  - Avoiding inconsistencies
- Thesauri often are constructed and maintained by trustworthy organisations
- Many thesauri cover a specific field of interest contain general terms but no enterprise-specific terms (product names, projects etc.) Adding them requires effort for maintenance.