

## 4. Wissensrepräsentation für Texte

Norbert Fuhr

# Problemstellung

## Repräsentation von Textinhalten:

Problem: Konzepte aus der Anfrage können im Text auf unterschiedlichste Weise formuliert werden

## Lösungsansätze

semantischer Ansatz Zuordnung von Deskriptionen zu Texten →  
Dokumentationssprachen

### Freitextsuche

informatischer Ansatz: Textretrieval als  
Zeichenkettensuche

computerlinguistischer Ansatz: i.w. Normalisierung  
von Wortformen

# Allgemeine Eigenschaften

formulierungsunabhängige Repräsentation von Textinhalten  
durch Verwendung eines speziellen Vokabulars

- ▶ Klassifikationen
- ▶ Thesauri
- ▶ RDF

# Klassifikationen

Strukturierung eines Wissensgebietes nach einem vorgegebenen formalen Schema

*z.B. Dezimalklassifikation: Baum der Ordnung 10*

Ein Dokument wird in der Regel einer oder wenigen Klassen zugeordnet

(ursprünglich für Bibliotheken entwickelt - ein Buch kann nur an einem Platz stehen!)

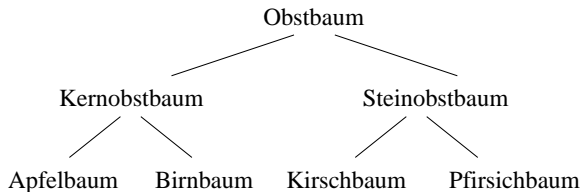
## Beispiele für Klassifikationen

- ▶ Web-Kataloge (z.B. *Yahoo!*)
- ▶ Klassifikationen in bestimmten Fachgebieten/Anwendungsbereiche:
  - LCC Library of Congress Classification
  - DDC Dewey Decimal Classification
  - UDC Universal Decimal Classification
  - MSc Mathematics Subject Classification
  - CCS ACM Computing Classification system

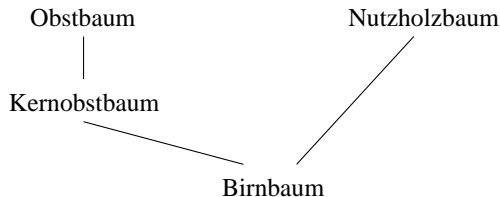
# Eigenschaften von Klassifikationssystemen

## Monohierarchie — Polyhierarchie

### Monohierarchie:



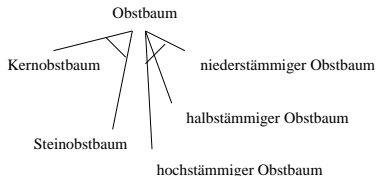
### Polyhierarchie:



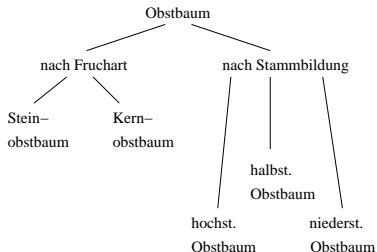
# Monodimensionalität — Polydimensionalität

Problem: auf einer Stufe gibt es mehrere Kriterien, nach denen eine weitere Aufteilung in Unterklassen vorgenommen werden kann

## Polydimensionalität



## aufgelöst



## Analytische vs. synthetische Klassifikation

analytische Klassifikation: top-down Vorgehensweise  
(wie oben)

synthetische Klassifikation: bottom-up

1. Erhebung der Merkmale der zu klassifizierenden Objekte und Zusammenstellung im Klassifikationssystem
2. Bildung der Klassen durch Kombination der Merkmale



# Facettenklassifikation

**Beispiel:** Facettenklassifikation Obstbäume

Facette	Facette	Facette
A Fruchtart	B Stammart	C Erntezeit
A1 Apfel	B1 hochstämmig	C1 früh
A2 Birne	B2 halbstämmig	C2 mittel
A3 Kirsche	B3 niederstämmig	C3 spät
A4 Pfirsich		
A5 Pflaume		

*A1B3C1 = niederstämmiger Frühapfelbaum*

**Regeln:**

- ▶ Facetten müssen disjunkt sein
- ▶ monodimensionale Unterteilung innerhalb einer Facette

# Yahoo! – main categories

## **Arts & Humanities**

Literature, Photography...

## **Business & Economy**

B2B, Finance, Shopping, Jobs...

## **Computers & Internet**

Internet, WWW, Software, Games...

## **Education**

College and University, K-12...

## **Entertainment**

Cool Links, Movies, Humor, Music...

## **Government**

Elections, Military, Law, Taxes...

## **Health**

Medicine, Diseases, Drugs, Fitness...

## **News & Media**

Full Coverage, Newspapers, TV...

## **Recreation & Sports**

Sports, Travel, Autos, Outdoors...

## **Reference**

Libraries, Dictionaries, Quotations...

## **Regional**

Countries, Regions, US States...

## **Science**

Animals, Astronomy, Engineering...

## **Social Science**

Archaeology, Economics, Languages...

## **Society & Culture**

People, Environment, Religion...

# Yahoo! – Computers & Internet

Art@	Employment@
Bibliographies (6)	Ethics (18)
Communications and Networking (1146)	Games@
Computer Science@	Graphics (316)
Contests (26)	Hardware (2355)
Conventions and Conferences@	History (106)
Countries, Cultures, and Groups (38)	Humor@
Cyberculture@	Industry Information@
Data Formats (485)	Internet (6066)
Desktop Customization@	Magazines@
Desktop Publishing (53)	Mobile Computing (65)
Dictionaries (24)	Multimedia (690)
	Music@
	News and Media (205)
	...

# Yahoo!

- ▶ Polyhierarchie
- ▶ Tiefe der Hierarchie variiert
- ▶ Dokumente können beliebigen Klassen zugeordnet werden

# ACM Computing Classification System

Ursprünglich Klassifikation in der Zeitschrift *ACM Computing Reviews*,  
wird vielfach als Standard-Klassifikation in der Informatik  
verwendet.

## Elemente:

- ▶ **general terms**: vorgegebene Menge von allgemeinen Begriffen
- ▶ **classification codes**: dreistufige monohierarchische Klassifikation
- ▶ **subject headings**: vorgegebene Menge von natürlichsprachlichen Bezeichnungen für jede einzelne Klasse, die diese weiter differenzieren; außerdem alle Eigennamen
- ▶ **free terms**: zusätzliche, frei wählbare Stichwörter

## General terms:

These apply to any elements of the tree that are relevant

ALGORITHMS	MANAGEMENT
DESIGN	MEASUREMENT
DOCUMENTATION	PERFORMANCE
ECONOMICS	RELIABILITY
EXPERIMENTATION	SECURITY
HUMAN FACTORS	STANDARDIZATION
LANGUAGES	THEORY
LEGAL ASPECTS	VERIFICATION

# Übersicht über die Hauptklassen

- A. GENERAL LITERATURE
- B. HARDWARE
- C. COMPUTER SYSTEMS ORGANIZATION
- D. SOFTWARE
- E. DATA
- F. THEORY OF COMPUTATION
- G. MATHEMATICS OF COMPUTING
- H. INFORMATION SYSTEMS
- I. COMPUTING METHODOLOGIES
- J. COMPUTER APPLICATIONS
- K. COMPUTING MILIEUX

## H.3 INFORMATION STORAGE AND RETRIEVAL

### H.3.0 General

### H.3.1 Content Analysis and Indexing

- Abstracting methods

- Dictionaries

- Indexing methods

- Linguistic processing

- Thesauruses

### H.3.2 Information Storage

- File organization

- Record classification

### H.3.3 Information Search and Retrieval

- Clustering

- Query formulation

- Retrieval models

- Search process

- Selection process

### H.3.4 System and Software...



# Eigenschaften der ACM-CCS

- ▶ Monohierarchie
- ▶ feste Tiefe (vier Ebenen)
  - ▶ Buchstaben/Ziffern-Code für Ebene 1–3
  - ▶ “subject heading” auf Ebene 4
- ▶ Dokumente können nur der 4. Ebene zugeordnet werden

# Dezimalklassifikation

Ursprung: Dewey Decimal Classification (DDC),  
1876 von Melvil Dewey (USA) entwickelt  
Universalklassifikation zur Aufstellung von Buchbeständen

Weiterentwickelt durch Paul Otlet und Henri Lafontaine (Belgien)  
zur **Universellen Dezimalklassifikation** (DK)  
(im Gegensatz zur DDC kaum noch benutzt)

## Grundelemente der DK:

- ▶ Hierarchisch gegliederten Klassen (130000)
- ▶ Anhängeszahlen zur Facettierung
- ▶ Sonderzeichen zur Verknüpfung mehrerer DK-Zahlen

# Hauptklassen

Die DK-Haupttafeln umfassen die Hauptabteilungen:

- 0 Allgemeines
- 1 Philosophie
- 2 Religion, Theologie
- 3 Sozialwissenschaften, Recht, Verwaltung
- 4 (zur Zeit nicht belegt)
- 5 Mathematik, Naturwissenschaften
- 6 Angewandte Wissenschaften, Medizin, Technik
- 7 Kunst, Kunstgewerbe, Photographie, Musik, Spiel, Sport
- 8 Sprachwissenschaft, Philologie, Schöne Literatur, Literaturwissenschaft
- 9 Heimatkunde, Geographie, Biographien, Geschichte

# Beispiel für die Untergliederung einer Hauptklasse

## Beispiel:

3 Sozialwissenschaften, Recht, Verwaltung

33 Volkswirtschaft

336 Finanzen. Bank- und Geldwesen

336.7 Geldwesen. Bankwesen. Börsenwesen

336.76 Börsenwesen. Geldmarkt. Kapitalmarkt

336.763 Wertpapiere. Effekten

336.763.3 Obligationen. Schuldverschreibungen

336.763.31 Allgemeines

336.763.311 Verzinsliche Schuldbriefe

336.763.311.1 Langfristig verzinsliche Schuldbriefe

# Facettierende Elemente

Anhängezahlen: durch spezielle Zeichen eingeleitet  
allgemeine Anhängezahlen: Facetten, die überall in der DK  
verwendet werden dürfen

Zeichenfolgen/Facetten:

= Sprache

(0...) Form

(...) Ort

(=...) Rassen und Völker

„...“ Zeit

.00 Gesichtspunkt

-05 Person

spezielle Anhängezahlen: nur für bestimmte Klassen innerhalb der  
DK erlaubt

# Verknüpfung von DK-Zahlen

spezielle Sonderzeichen zur Verknüpfung von DK-Zahlen:

- + Aufzählung mehrerer Sachverhalte
  - : symmetrische Beziehung zwischen zwei Sachverhalten (umkehrbar)
  - :: asymmetrische Beziehung zwischen zwei Sachverhalten
- / Erstreckungszeichen (zur Zusammenfassung mehrerer nebeneinanderstehender DK-Zahlen)
- ' Zusammenfassungszeichen zur Bildung neuer Sachverhalte aus der Kombination einzelner DK-Komponenten

# Thesauri

DIN 1463:

„Thesaurus ist geordnete Zusammenstellung von Begriffen mit ihren (natürlichsprachlichen) Bezeichnungen.

Merkmale eines Thesaurus:

- a) terminologische Kontrolle durch
  - ▶ Erfassung von Synonymen
  - ▶ Kennzeichnung von Homonymen und Polysemen
  - ▶ Festlegung von Vorzugsbenennungen
- b) Darstellung von Beziehungen zwischen Begriffen“

# Terminologische Kontrolle

Reduktion von Mehrdeutigkeiten und Unschärfe der natürlichen Sprache

## Synonymkontrolle

Zusammenfassung von Bezeichnungen zu Äquivalenzklassen

Arten von Synonymie:

- ▶ Schreibweisenvarianten

*Friseur — Frisör*

*UN — UNO — Vereinte Nationen*

- ▶ unterschiedlichen Konnotationen, Sprachstile, Verbreitung

*Telefon — Fernsprecher*

*Pferd — Gaul*

*Myopie — Kurzsichtigkeit*

- ▶ Quasi-Synonyme

*Schauspiel — Theaterstück*

*Rundfunk — Hörfunk*



## Synonymkontrolle

Im Thesaurus werden darüber hinaus Begriffe mit geringen / irrelevanten Bedeutungsunterschieden zu Äquivalenzklassen zusammengefaßt:

- ▶ unterschiedliche Spezifität  
*Sprachwissenschaft — Linguistik*
- ▶ Antonyme  
*Härte — Weichheit*
- ▶ zu spezieller Unterbegriff  
*Weizen — Winterweizen*
- ▶ Gleichsetzung von Verb und Substantiv / Tätigkeit und Ergebnis  
*Wohnen — Wohnung*

## Polysemkontrolle

Aufteilung von einer (mehrdeutigen) Bezeichnung auf mehrere Äquivalenzklassen

- ▶ Homonyme (*Bs. Tenor, Vollzug, modern, Hochzeit*)
- ▶ Polyseme (*Bs. Bank, Deckung*)

## Zerlegungskontrolle

Problem: Wie spezifisch sollen einzelne Begriffe im Thesaurus sein?  
*„Donaudampfschiffahrtskapitän“*

Nachteile zu spezieller Begriffe:

- ▶ Thesaurus zu umfangreich / unübersichtlich
- ▶ nur wenige Dokumente zu einer Äquivalenzklasse

### UNITERM-Verfahren:

Nur Begriffe, die nicht weiter zerlegbar sind (Unitersms)  
Verkettung von Unitersms zur Wiedergabe eines Sachverhaltes  
(Postkoordination)

Nachteil: größere Unschärfe beim Retrieval

*Baum + Stamm = Baumstamm / Stammbaum*

Thesaurusmethode: Kompromiß zwischen beiden Ansätzen

# Äquivalenzklasse — Deskriptor

Terminologische Kontrolle liefert Äquivalenzklassen von Bezeichnungen

Darstellung dieser Äquivalenzklassen:

- ▶ Thesaurus ohne Vorzugsbenennung:  
Gleichbehandlung aller Elemente der Äquivalenzklasse
- ▶ Thesaurus mit Vorzugsbenennung:  
Auswahl eines Elementes der Äquivalenzklasse zur Benennung  
=Deskriptor  
(im folgenden nur Thesauri mit Vorzugsbenennung betrachtet)

# Beziehungsgefüge des Thesaurus

## Äquivalenzrelation

zwischen Nicht-Deskriptoren und Deskriptoren

Bezeichnungen:

**BS** Benutze Synonym (use)

**BF** Benutzt für (used for, UF)

*Fernsprecher* **BS** *Telefon*

*Telefon* **BF** *Fernsprecher*

## Hierarchische Relation

zwischen Deskriptoren

Bezeichnungen:

**UB** Unterbegriff (narrower term, NT)

**OB** Oberbegriff (broader term, BT)

*Obstbaum* **UB** *Steinobstbaum*

*Steinobstbaum* **OB** *Obstbaum*

## Assoziationsrelation

zwischen begriffsverwandten Deskriptoren, symmetrisch

Bezeichnung: **VB** verwandter Begriff (see also, SEE)

*Obstbaum* **VB** *Obst*

*Obst* **VB** *Obstbaum*

# Darstellung des Thesaurus

## Deskriptor-Einträge

- ▶ Begriffsnummer
- ▶ Notation / Deskriptor-Klassifikation
- ▶ Scope note / Definition
- ▶ Synonyme
- ▶ Oberbegriffe / Unterbegriffe
- ▶ Verwandte Begriffe
- ▶ Einführungs-/Streichungsdatum



# Gesamtstruktur des Thesaurus

(in gedruckter Form)

Hauptteil mit den Deskriptor-Einträgen  
alphabetisch / systematisch geordnet

zusätzliche Register mit Verweisen auf die Deskriptor-Einträge

- ▶ systematisch / alphabetisch (komplementär zum Hauptteil)
- ▶ Index für Komponenten mehrgliedriger Bezeichnungen  
KWIC — keyword in context  
KWOC — keyword out of context

# Beispiel: INSPEC-Thesaurus

<b>Information retrieval</b>		<b>Query processing</b>	
<i>UF</i>	CD-ROM searching Data access Document retrieval Online literature searching	<i>UF</i>	Data querying Database querying Query optimisation
	Retrieval, information	<i>BT</i>	Information retrieval
<i>BT</i>	Information science	<i>RT</i>	Database management systems Database theory DATALOG Query languages
<i>NT</i>	Query formulation Query processing Relevance feedback	<b>Query formulation</b>	
<i>RT</i>	Bibliographic systems Information analysis Information storage Query languages	<i>UF</i>	Search strategies
		<i>BT</i>	Information retrieval
		<b>Relevance feedback</b>	
		<i>BT</i>	Information retrieval

# Beispiel-Thesaurus

0.0058

Magnetband

VB Magnetbandlaufwerk

0,0045

Magnetbandgerät

BS Magnetbandlaufwerk NE7

0. 0046

Magnetbandkassette

NO NE83

BF Kassette

BF MB-Kassette

OB Datenträger

VB Magnetbandkassettenlaufwerk

0.0051

Magnetbandkassettengerät

BS Magnetbandkassettenlaufwerk NE7

Magnetismus (Forts.)

BF Halleffekt

BF Induktion

OB Elektrodynamik

UB Magnetfeld

BIK Geophysik

BFK Erdmagnetismus

BIK Optik

BFK Faraday-Effekt

0.0070

Magnetkarte

NO NE87

BF Telefonkärtchen

OB Datenträger

VB Kartensystem

0.0073

Magnetkartensystem

NO ECS

OB Kartensystem

0.0050  
Magnetbandkassettenlaufwerk  
NO NE7  
BF Magnetbandkassettengerät  
BF MB-Kassettengerät  
OB Datenausgabegerät  
OB Dateneingabegerät  
OB Datenspeichertechnik  
VB Magnetbandkassette

0.0044  
Magnetbandlaufwerk  
NO NE7  
BF Magnetbandgerät  
OB Bandgerät  
OB Datenausgabegerät  
OB Dateneingabegerät  
OB Datenspeichertechnik  
VB Magnetband

0.0059  
Magnetfeld  
NO WD2  
OB Magnetismus

0.0060  
Magnetismus  
NO WD2  
BF Barkhausen-Effekt  
BF Ferromagnetismus

0.0074  
Magnetkartentelefon  
NO GK72  
BF Makatel  
OB Kartentelefon

0 0077  
Magnetplatte  
NO NE82  
OB Datenspeicher  
OB Datenträger  
VB Magnetplattenlaufwerk  
BIK Datenspeicher  
BFK Plattenspeicher

0.0081  
Magnetplattengerät  
BS Magnetplattenlaufwerk

0.0079  
Magnetplattenlaufwerk  
NO NE7  
BF Magnetplattengerät  
OB Datenausgabegerät  
OB Dateneingabegerät  
OB Datenspeichertechnik  
VB Magnetplatte

# Thesauruspflege

Anpassung des Thesaurus an Veränderungen in der Anwendung  
notwendig aufgrund von

- ▶ Entwicklung des Fachgebietes  
*objektorientierte Datenbanken, multimediale Systeme*
- ▶ Entwicklung der Fachsprache
- ▶ Indexierungsverhalten / Indexierungsergebnisse
- ▶ Benutzerverhalten
- ▶ Rechercheergebnisse

Problem: Überwachung der Konsistenz des Thesaurus

# RDF

(Resource Description Framework)  
vom W3C im Rahmen der 'Semantic Web'-Initiative geförderte  
Beschreibungssprache

Idee: ausdrucksstärkere Beschreibungssprache

- ▶ Instanzen zu Konzepten
- ▶ beliebige Beziehungen zwischen Instanzen ausdrücken
- ▶ Statements der Art Subjekt-Prädikat-Objekt

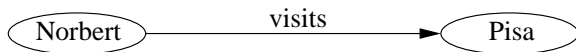
## RDF: basic concepts

**Resource** object on the WWW, e.g. Web page, database  
naming of resources: Uniform Resource Identifier (URI)

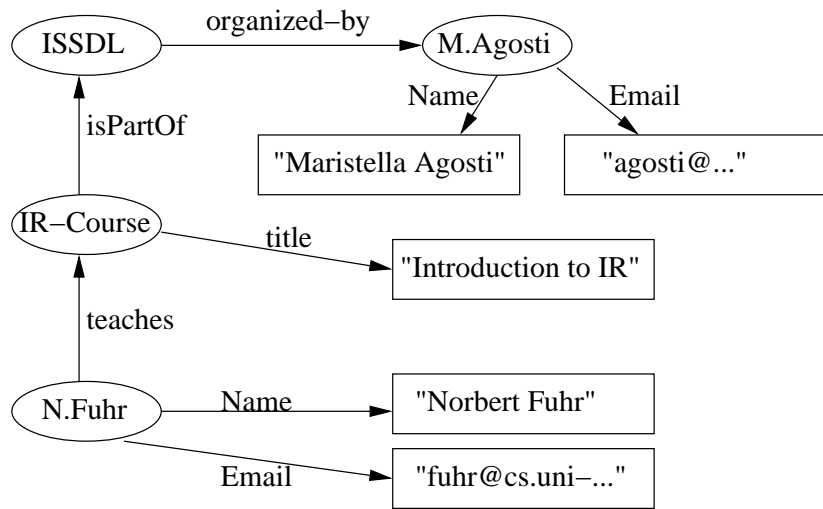
**Literal** special type of resource, with string value, no explicit URI

**Property** aspect / attribute / characteristics / relation

**Statement** resource + named property + value of property  
(subject, predicate, object)



## RDF example





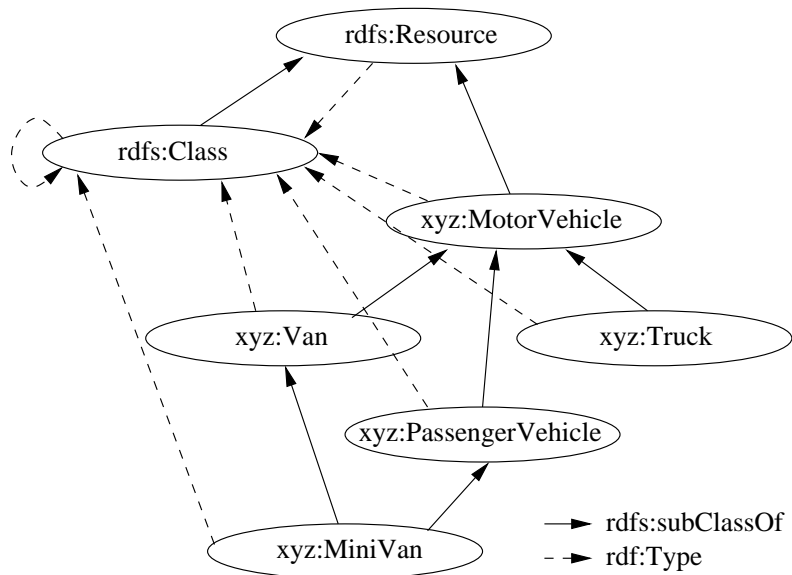
# RDF schemas

similar to semantic networks / description logics

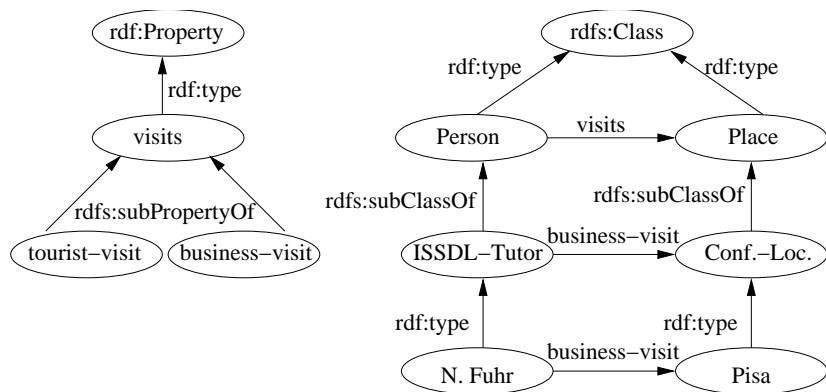
describes relationships between types of resources and/or properties

- ▶ fundamental concepts
  - ▶ rdfs:Resource
  - ▶ rdf:Property
  - ▶ rdfs:Class
- ▶ schema definition concepts
  - ▶ rdf:type
  - ▶ rdfs:subClassOf
  - ▶ rdfs:subPropertyOf
  - ▶ rdfs:seeAlso
  - ▶ rdfs:isDefinedBy

## RDFS example: resource hierarchy



## RDFS example: resource and property hierarchies



# Freitextsuche

Voraussetzungen:

- ▶ Zerlegung von Texten in Wörter
- ▶ (Stopworteliminierung)
- ▶ (Satzendeerkennung)

## Probleme bei der Freitextsuche:

- ▶ **Homographen**

*Tenor: Sänger / Ausdrucksweise*

- ▶ **Polyseme**

*Bank: Sitzgelegenheit / Geldinstitut*

- ▶ **Flexionsformen**

*Haus – (des) Hauses – Häuser  
schreiben – schreibt – schrieb – geschrieben*

- ▶ **Derivationsformen**

*Formatierung – Format – formatieren*

- ▶ **Komposita** (zusammengesetzte Wörter)

*Donaudampfschiffahrtsgesellschaftskapitän Bundeskanzlerwahl*

- ▶ **Nominalphrasen** (aus mehreren Nomen zusammengesetzte Begriffe)

*Wahl des Bundeskanzlers  
information retrieval – retrieval of information – information  
was retrieved*

Das Problem der Wortwahl bleibt ungelöst!

# Informatischer Ansatz

## Zeichenketten-Operatoren für die Freitextsuche

### ▶ Truncation

Front-/End-Truncation,

beschränkt (\$) / unbeschränkt(#)

*schreib#*: schreiben, schreibt, schreibst, schreibe

*schreib\$\$*: schreiben, schreibst

*#schreiben*: schreiben, beschreiben, anschreiben, verschreiben

*\$\$schreiben*: beschreiben, anschreiben

### ▶ (Mitten-)Maskierung

*do\$umentation*: documentation, Dokumentation

*schr\$\$b#*: schreiben, schrieb / schrauben

Truncation und Maskierung dienen dazu, Flexions- und Derivationsformen von Wörtern zusammenzuführen (auch für Komposita geeignet)

**Vorteil:** weniger Schreibarbeit als beim expliziten Aufzählen

**Nachteil:** möglicherweise unerwünschte Wörter dabei

## Kontextoperatoren

zur Suche nach mehrgliedrigen Ausdrücken (Nominalphrasen)

*information AND retrieval:*

boolesche Operatoren beziehen sich nur auf das Vorkommen irgendwo im Text!

- ▶ genauer Wortabstand (\$):  
*retrieval \$ information: retrieval of information, retrieval with information loss*
- ▶ maximaler Wortabstand (#):  
*text # # retrieval: text retrieval, text and fact retrieval*
- ▶ Wortreihenfolge (,):  
*information # , retrieval: information retrieval, retrieval of information*
- ▶ gleicher Satz (.):  
*information # retrieval.* matcht nicht  
*... this information. Retrieval of data ...*  
aber auch nicht:  
*... storage of information. Its retrieval ...*



# Computerlinguistischer Ansatz

Arten von Verfahren:

**graphematische Verfahren** auf der Analyse von Buchstabenfolgen basierende Algorithmen, hauptsächlich zur Zusammenführung von Flexions- oder Derivationsformen (Morphologie)

**lexikalische Verfahren** Wörterbuch-basierte Verfahren zur Zusammenführung von Flexions- oder Derivationsformen sowie von mehrgliedrigen Ausdrücken

**syntaktische Verfahren** zur Identifikation von mehrgliedrigen Ausdrücken

# Graphematische Verfahren

(für die englische Sprache)

- ▶ Grundformreduktion

Zurückführen auf die Grundform, d.h.

Substantive im Nominativ Singular,

Verben im Infinitiv

- ▶ lexikographische Grundform

entsteht durch Abtrennen der Flexionsendung und ggfs.

Rekodierung

*applies* → *appl* → *apply*

- ▶ formale Grundform

nur Abtrennen von Endungen, ohne Rekodierung

*activities* → *activit*

- ▶ Stammformreduktion

Entfernen der Derivationsendungen, d.h. Zurückführen auf den

Wortstamm

*computer, compute, computation, computerization* → *comput*

# Lexikographische Grundformreduktion

(nach Kuhlen 77)

% alle Vokale (einschließlich Y)

\* alle Konsonanten

/ ,oder'

⊞ Leerzeichen

→ ,zu'

← ,aus'

¬ ,nicht'

- 1) **IES** → **Y**
- 2) **ES** → ~~**ß**~~      wenn \*O / CH / SH / SS / ZZ / X  
vorangehen
- 3) **S** → ~~**ß**~~      wenn \* / E / %Y / %O / OA / EA vorangehen
- 4) **S'** → ~~**ß**~~  
**IES'** → **Y**  
**ES'** → ~~**ß**~~
- 5) **'S** → ~~**ß**~~  
**'** → ~~**ß**~~
- 6) **ING** → ~~**ß**~~      wenn \*\* / % / X vorausgehen  
**ING** → **E**      wenn %\* vorausgehen
- 7) **IED** → **Y**
- 8) **ED** → ~~**ß**~~      wenn \*\* / % / X vorausgehen  
**ED** → **E**      wenn %\* vorausgehen

---

Regel 1 **IES** → **Y**

---

*Beispiele zu 1:*

APPLIES → APPLY

IDENTIFIES → IDENTIFY

ACTIVITIES → ACTIVITY

---

Regel 2 **ES** → **ɛ**, wenn \*O / CH / SH / SS / ZZ /  
X vorangehen

---

*Beispiele zu 2:*

BREACHES	→	BREACH
PROCESSES	→	PROCESS
FISHES	→	FISH
COMPLEXES	→	COMPLEX
TANGOES	→	TANGO
BUZZES	→	BUZZ

---

Regel 3 **S** → **Ø**, wenn \* / E / %Y / %O / OA /  
EA vorangehen

---

*Beispiele zu 3:*

METHODS	→	METHOD
HOUSES	→	HOUSE
BOYS	→	BOY
RADIOS	→	RADIO
COCOAS	→	COCOA
FLEAS	→	FLEA

---

Regel 4	<b>S'</b>	→	<b>Ø</b>
	<b>IES'</b>	→	<b>Y</b>
	<b>ES'</b>	→	<b>Ø</b>

---

*Beispiele zu 4:*

MOTHERS'	→	MOTHER
LADIES'	→	LADY
FLAMINGOES	→	FLAMINGO



---

Regel 5 'S → ~~B~~  
' → ~~B~~

---

*Beispiele zu 5:*

MOTHER'S → MOTHER  
CHILDREN'S → CHILDREN  
PETRUS' → PETRUS

---

Regel 6   **ING** → ~~B~~, wenn \*\* / % / X vorausgehen  
              **ING** → **E**, wenn %\* vorausgehen

---

*Beispiele zu 6:*

DISGUSTING → DISGUST  
GOING → GO  
MIXING → MIX  
LOOSING → LOOSE  
RETRIEVING → RETRIEVE

---

Regel 7   **IED** → **Y**

---

*Beispiel zu 7:*

SATISFIED → SATISFY

---

Regel 8   **ED** → **B**, wenn \*\* / % / X vorausgehen

**ED** → **E**, wenn %\* vorausgehen

---

*Beispiel zu 8:*

DISGUSTED → DISGUST

OBEYED → OBEY

MIXED → MIX

BELIEVED → BELIEVE

# Lexikalische Verfahren

besonders für stark flektierte Sprachen (z.B. deutsch) geeignet

Relationen im Wörterbuch:

- ▶ Flexionsform (Vollformen) — zugehörige Grundform  
*Hauses - Haus, ging - gehen*
- ▶ Derivationsform — zugehörige Grundformen  
*Lieblosigkeit — lieblos, Berechnung — rechnen*
- ▶ Komposita — zugehörige Dekomposition  
*Haustür — Tür, Armbanduhr — Uhr.*

# Syntaktische Verfahren

(zur Behandlung von Nominalphrasen)

1. Wortklassenbestimmung
2. Parsing
3. Identifikation von Nominalphrasen
4. Head-Modifier-Strukturen
5. Matching

# Wortklassenbestimmung

typische Wortklassen:

AT	article	PN	personal pronoun
BEZ	“is”	RB	adverb
CONJ	conjunction	RBR	comparative adverb
IN	preposition	TO	“to”
JJ	adjective	VB	verb, base form
JJR	comparative adjective	VBD	verb, past tense
MD	modal (can, have, may,...)	VBG	verb, present participle, gerund
NN	singular or mass noun	VBN	verb, past participle
NNP	singular proper noun	VBP	verb, non 3rd singular present
NNS	plural noun	VBZ	verb, 3rd singular present
PERIOD	.:?!)	WDT	<i>wh</i> -determiner (what, which)

## Datenquellen für die Wortklassenbestimmung

- ▶ (Vollformen-)Wörterbuch
- ▶ graphematische Verfahren  
(insbesondere für nicht im Wörterbuch enthaltene Wörter)

### Problem:

Wortklassenbestimmung in wenig flektierten Sprachen

*The boys **play** football vs.*

*She saw the new **play***

→ zusätzliche Berücksichtigung der syntaktischen Struktur  
(Bigramme, Trigramme) notwendig:

AT NNS VBP/NN NN → VBP

PN VBD AT JJ NN/VBP → NN

## Beispiel:

Wortklassenbestimmung basierend auf dem Kuhlen-Algorithmus

Nr.	Regel	Klasse
1	<b>IES</b> → <b>Y</b>	NNS/VBP
2	<b>ES</b> → <del><b>B</b></del>	NNS/VBP
3	<b>S</b> → <del><b>B</b></del>	NNS/VBP
4	<b>S'</b> → <del><b>B</b></del> , <b>IES'</b> → <b>Y</b> , <b>ES'</b> → <del><b>B</b></del>	NNS
5	<b>'S</b> → <del><b>B</b></del> , <b>'</b> → <del><b>B</b></del>	NN
6	<b>ING</b> → <del><b>B</b></del> , <b>ING</b> → <b>E</b>	VBG
7	<b>IED</b> → <b>Y</b>	VBD/VBN/JJ
8	<b>ED</b> → <del><b>B</b></del> , <b>ED</b> → <b>E</b>	VBD/VBN/JJ



## Kombination von graphematischem Verfahren und Vollformen-WB

(Rückführung unbekannter Wörter auf WB-Einträge mittels graphematischer Verfahren) [Mikheev 98]

Präfix	WB-Klassen	Wortklassen
re	JJ NN VBG	JJ NN VBG
ex	NN	NN
self-	NN	NN
inter	JJ	JJ
non	JJ	JJ
un	RB	RB
dis	JJ	JJ
anti-	NN	JJ
de	JJ VBD VBN	JJ VBD VBN
in	RB	RB

Postfix	WB-Klassen	Wortklassen
ment	NN VB VBP	NN
ing	NN VB VBP	JJ NN VBG
ed	NN VB VBP	JJ VBD VBN
s	NN VB VBP	NNS VBZ
ly	JJ NN RB	RB
ness	JJ	NN
ship	NN	NN
able	NN VB VBP	JJ
s	NN	NNS

## Wortklassenbestimmung mit deterministischem Tagger

nur 70 % korrekte Zuordnungen! [Greene & Rubin 77]

aber:

die meisten Wörter kommen in einer bevorzugten Wortklasse vor

to **flour** a pan

to **web** the final report

→ seltene Verwendungen ignorieren!

90 % korrekte Zuordnungen! [Charniak et al. 93]

Verbesserung:

statistische Ansätze zur Berücksichtigung der syntaktischen Struktur (z.B. Markov-Modelle)

liefern 95...97 % korrekte Zuordnungen

# Parsing

eine einfache Grammatik:

S → NP VP  
NP → AT? JJ\* NNS+  
→ AT? JJ\* NN+  
→ NP PP  
VP → VB PP  
→ VBZ  
→ VBZ NP  
PP → IN NP

Beispiele:

*The analysis of 25 indexing algorithms shows consistent retrieval performance.*

AT NN IN JJ NN NNS VBZ JJ NN NN

*A good indexing technique for Web retrieval is manual classification.*

AT JJ NN NN IN NN NN VBZ JJ NN

# Identifikation von Nominalphrasen

nur bestimmte syntaktische Strukturen relevant

→ partielles parsing

einfache Muster (ohne Unterscheidung (NN/NNP/NNS)):

phrase → NN NN+

→ NN+ IN JJ\* NN+

Beispiele:

*indexing algorithms*

*retrieval performance*

*retrieval of Web documents*

*retrieval of new documents*

# Head-Modifier-Strukturen

basierend auf der Transformation der Nominalphrasen in head-modifier-Strukturen:

**head:** Nomen, das die wesentliche Bedeutung der Nominalphrase ausdrückt:  
information *retrieval*, indexing *algorithm*

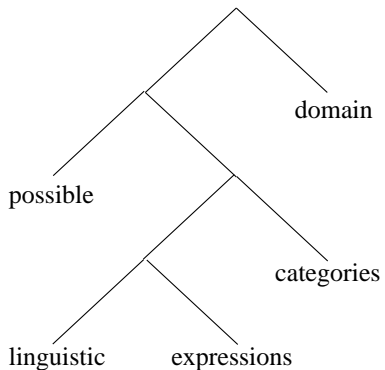
**modifier:** modifiziert/spezialisiert die Bedeutung des Heads

head-Modifier-Strukturen können geschachtelt auftreten

→ Darstellung als geschachtelte Listen oder Bäume

*((((multimedia, document), retrieval), system)*

the domain of possible categories of linguistic expressions



# Matching-Prozess

1. Nominalphrase in Head-Modifier-Struktur überführen  
(Transformationsregeln basierend auf der syntaktischen Struktur, ggfs. auch mit lexikalischen Angaben)
2. Vergleich: Head- bzw. Modifier-Rolle (bzgl. der gemeinsamen Wurzel) müssen übereinstimmen  
(einzelnes Nomen = Head)

## Beispiel:

*((semistructured, data), retrieval) system*) wird gematcht von

- ▶ *(retrieval, system)*
- ▶ *(semistructured, data)*
- ▶ *(data, retrieval)*

aber nicht von

- ▶ *(retrieval, data)*



# Beurteilung der Verfahren zur Repräsentation von Textinhalten

- ▶ Dokumentations-sprachen bieten prinzipiell Vorteile gegenüber der Freitextsuche  
aber: dieser Vorteil ist bislang experimentell nicht belegt, es gibt sogar gegenteilige Ergebnisse
- ▶ Erfahrungen aus TREC:  
halb-formale Konzepte (wie geographische und Datumsangaben) sind durch Freitextsuche nicht abzudecken
- ▶ wissensbasiertes IR:  
benötigt zunächst große Wissensbasen, die bislang nicht verfügbar sind (CYC-Project, semantic Web)
- ▶ syntaktische Verfahren:  
für Nominalphrasen
- ▶ maschinenlesbare Wörterbücher:  
für Nominalphrasen und zur Disambiguierung

# Einfache statistische Modelle

## Beispiel für computerlinguistischen Ansatz

### Text:

Experiments with Indexing Methods.

The analysis of 25 indexing algorithms has not produced consistent retrieval performance. The best indexing technique for retrieving documents is not known.

### Stoppworteliminierung:

experiments indexing methods analysis indexing algorithms  
produced consistent retrieval performance best indexing technique  
retrieving documents known

### Stammformreduktion:

experiment index method analys index algorithm produc consistent  
retriev perform best index techni retriev document

## „semantische“ Sicht

- ▶ Multimenge von Terms
- ▶ Formen des Vorkommens  
(Ort, Sicherheit)

### Modell:

- ▶ Abbildung auf Attribute
- ▶ Semantik durch Statistik!

Computerlinguistische Verfahren sind präziser (und benutzerfreundlicher) als der informatische Ansatz

*aber:*

alle Verfahren sind mit Fehlern behaftet!