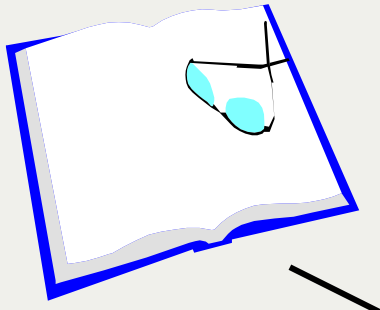
# Automatisches Indexieren

**Wörter - Texte - Information**

**Möglichkeiten und Grenzen automatischer Erschließungsverfahren**

# Indexieren und Automatisches Indexieren

**Dokumente**



Volltexte

bibliografische Referenzdaten

Zuteilung!

Abstract

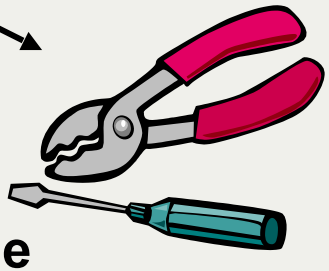
Thesaurus

Klassifikation

Schlagwort

**Inhaltserschließung**
**Indexierung**

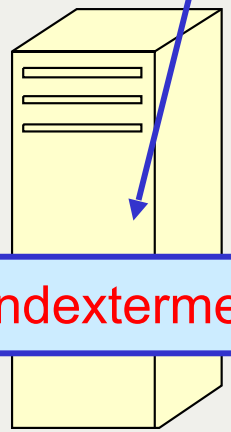
Repräsentationen des Dokumentinhalts



**Automatische Indexierung**

Verfahren zur Erzeugung von „guten“ Indextermen aus Dokumenten

Indexterme



**Dokumentenspeicher Suchsystem**

Linguistik

Statistik

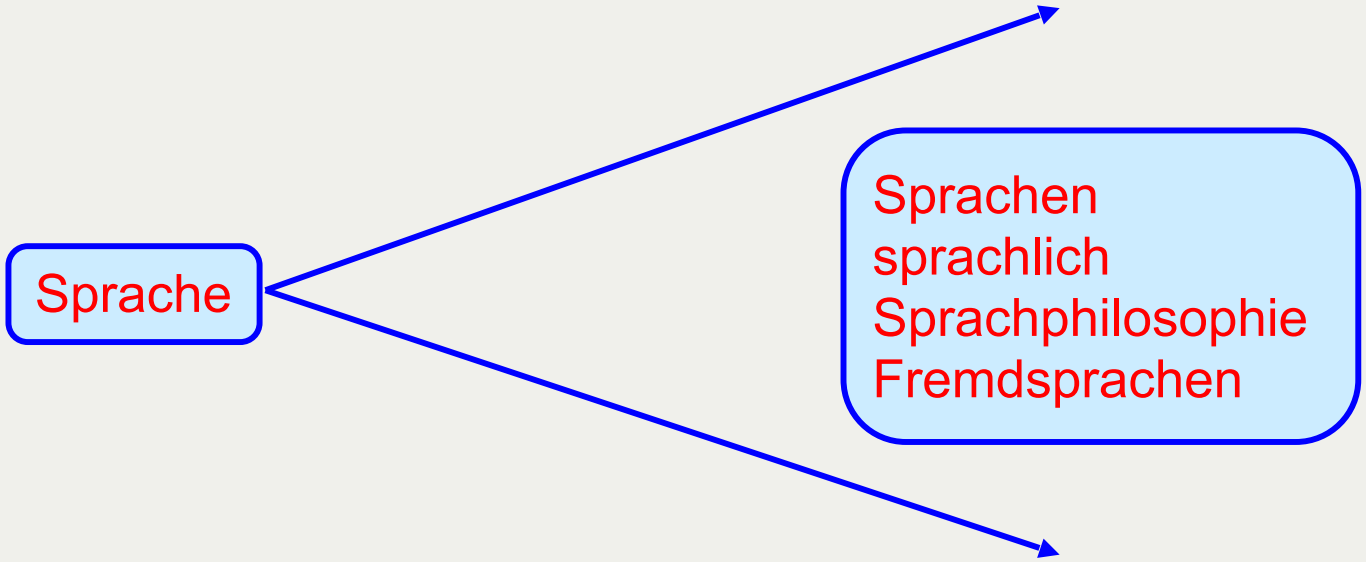
Extraktion!

# Verfahren zur Ermittlung „guter“ Indexterme

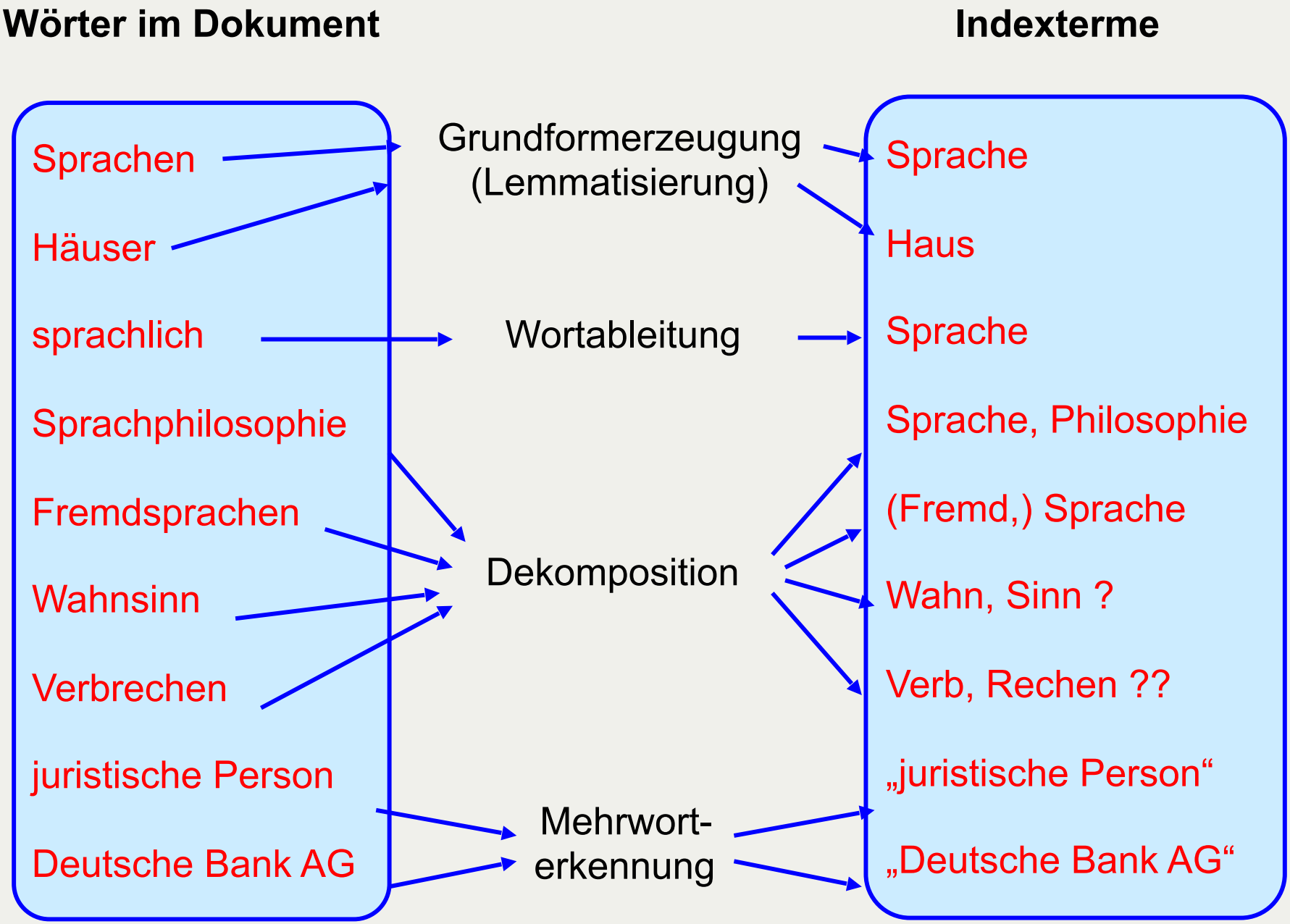
Die linguistische Hypothese:

„Wörter in Dokumenten sind selten gute Indexterme, weil sie sprachlich zu verschieden sind.“

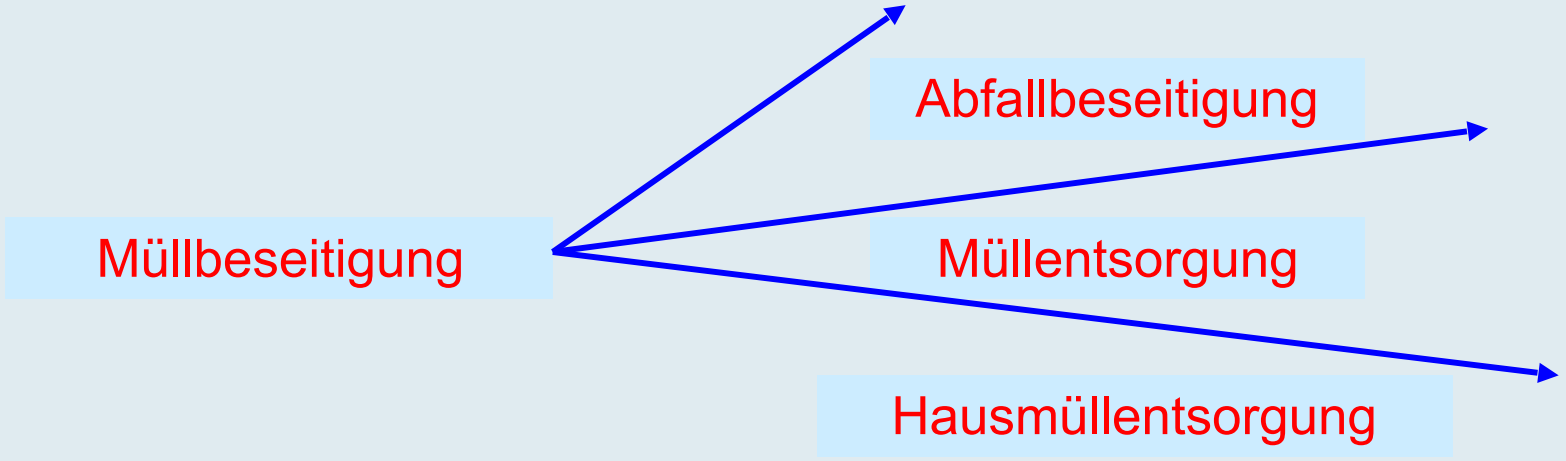
## Problem:
## Verschiedenartigkeit von Dokument- und Suchsprache



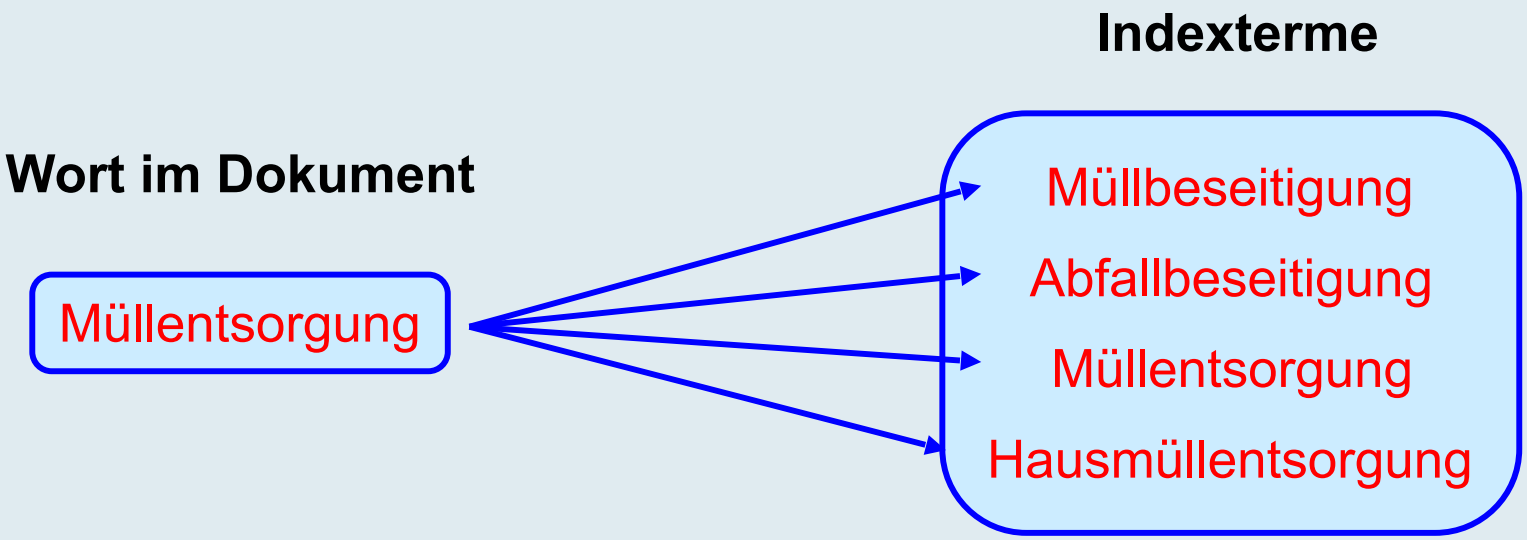
# Lösung: Einsatz einer morphologischen Komponente



## Problem: Suche im semantischen Umfeld



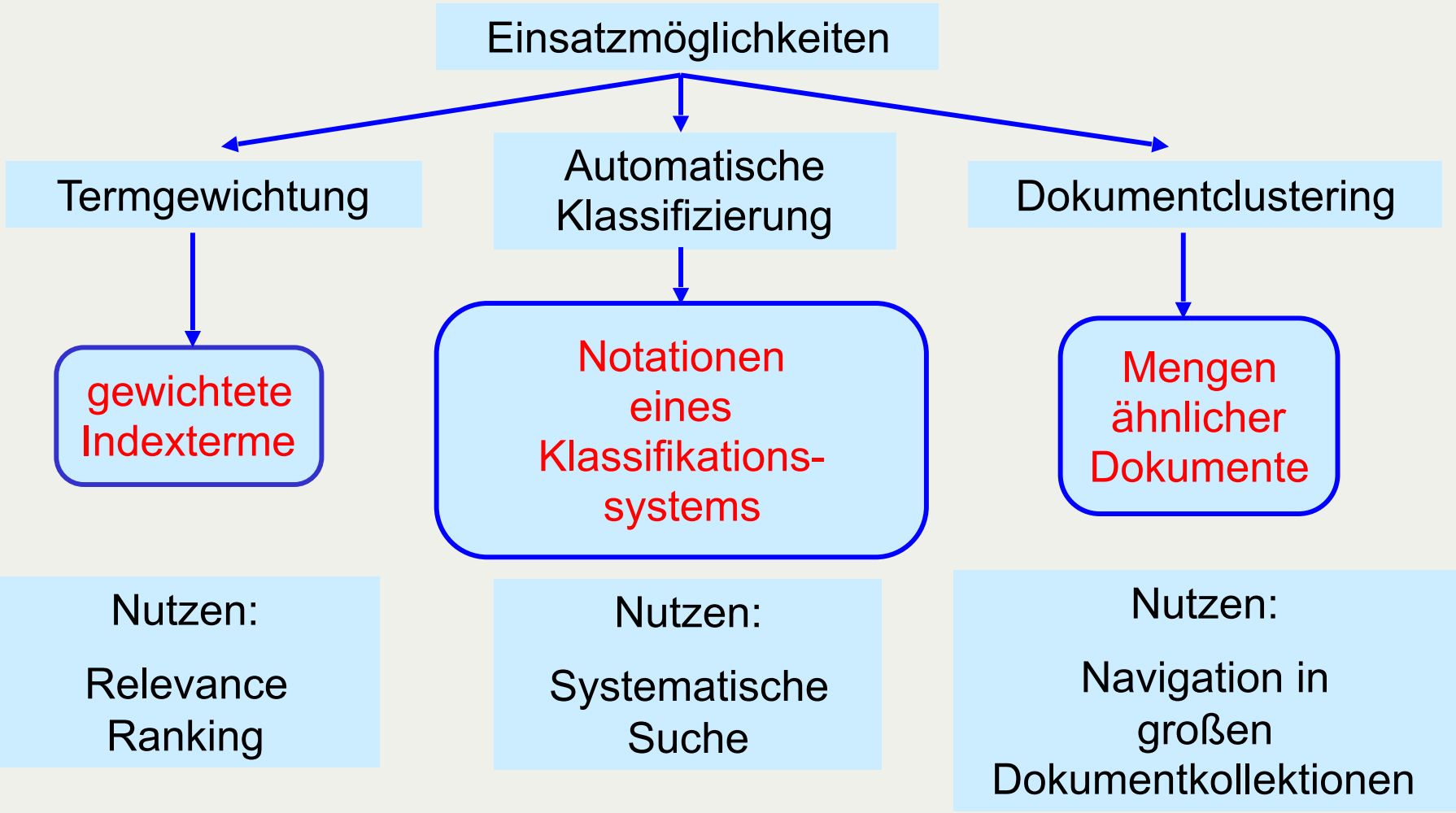

## Lösung: Einbindung von Synonym- ggf. hierarchischen Relationen



Die statistische Hypothese:

„Nur Wörter mit bestimmten statistischen Merkmalen (Häufigkeitsmerkmalen) sind „gute“ Indexterme.“

Verfahren: Ermittlung von Worthäufigkeiten (z.B. TF, IDF)



# Automatische Indexierung vs. Intellektuelle Erschließung

| | | |
|---|---|---|
| **Ziele** | • Retrievalverbesserung durch geeignete Indexterme<br>• Recallerhöhung auf Stichwortbasis | • einheitliche thematische Suche durch thematische Zusammenführung<br>• inhaltliche Repräsentation |
| **Aufwand** | • gering | • hoch |
| **Kosten** | • gering | • hoch |
| **Ergebnisse** | • Indexterme<br>• vorhersagbar<br>• beliebig reproduzierbar<br>• durch wiederholte Indexierungen leicht zu verbessern | • Schlagwörter, Deskriptoren, Notationen<br>• menschliche Fehlerquote<br>• endgültig<br>• inhaltlich verlässlich |
| **Qualität** | • abhängig von Qualität der Stichwörter im Quelltext | • hoch, bei konsistenter Erschließung |
| **Nutzen** | • für alle textbasierten Quelldokumente (auch erschlossene!) hoch<br>• insb. für nicht erschlossene und nicht zu erschließende Dokumente hoch | • bezogen auf die erschlossenen Dokumente hoch<br>• in Kollektionen mit niedriger Erschließungsquote gering |

# Was tun?

1. Linguistische Verfahren verbessern die Suchbarkeit allein auf der Basis von Stichwörtern. Aus nicht geeigneten bzw. nicht vorhandenen Stichwörtern lässt sich kein brauchbares Indexat machen!

2. Statistische Verfahren bewerten die vorhandene Stichwortbasis – s.o.!

3. Klassische Erschließungsziele wie Zusammenführung von Gleichem, vollständiger Nachweis, zuverlässige und einheitliche inhaltliche Suche sind nur durch Erschließung, nicht durch automatische Indexierung zu erreichen.

4. Eine Entscheidung für automatische Indexierung muss das Wissen um deren Leistungsfähigkeit insb. deren Grenzen der Leistungsfähigkeit berücksichtigen.

5. Einer Entscheidung für automatische Indexierung sollte daher eine Zielbestimmung hinsichtlich der gewünschten Retrievalmöglichkeiten jetzt und in 10 Jahren vorausgehen. Entscheidungen gegen eine Erschließung sind nur mit erheblichen Konsistenzverlusten umkehrbar.

**Allgemein sind die Grenzen automatischer Erschließungsverfahren dort erreicht, wo die Intelligenz beginnt.**

## Automatische Indexierung

## Linguistisch basierte Verfahren

| 2078 | » search |
|---|---|
| 1 | » search-aid |
| 1 | » search-and |
| 2 | » search-and-retrieval |
| 1 | » Search-Bots |
| 3 | » search-engine |
| 1 | » Search-Engine-Marketing |
| 1 | » Search-Engine-Optimierungen |
| 4 | » search-engines |
| 1 | » search-key |
| 1 | » search-off |
| 1 | » search-oriented |
| 1 | » search-outputs |
| 1 | » Search-process |
| 1 | » Search-Recherchesystem |
| 1 | » search-result |
| 1 | » search-Retrieve-Anweisung |
| 1 | » search-strategy |
| 1 | » search-support |
| 1 | » search-term |
| 1 | » search-topic |
| 5 | » searchability |
| 62 | » searchable |
| 2 | » SearchBank |
| 2 | » searchbots |
| 116 | » searched |
| 1 | » searchedu |
| 1 | » Searchengine |
| 2 | » Searchengines |
| 2 | » Searchenginewatch |
| 120 | » searcher |
| 1 | » SearchERIC |
| 214 | » searchers |
| 1 | » searchersystem |
| 485 | » searches |
| 1 | » searches-both |
| 1518 | » searching |
| 1 | » searching-and-indexing |
| 1 | » SearchMAGIC |
| 1 | » searchmethods |
| 1 | » SearchPad |
| 1 | » searchRetrieve |
| 2 | » searchs |
| 1 | » SearchSolver |
| 1 | » searchw |

## Zielvorstellung:

Abbildung aller Varianten von Indextermen auf einen gemeinsamen Indexterm

**„search“**

Verfahren

**sprachlich** begründete Reduzierung von

**Wortformen**

auf

**Grundformen**

oder

**Wortstämme**

Wörter in Texten „suchen“

mögliche Lexikoneinträge (Lemmata) „suche“

um Wortbildungs-elemente reduzierte Grundform „such“

## Bestandteile der Sprache I

- **Phonem** = kleinstes bedeutungsunterscheidendes Lautmerkmal

  **M**aus - **H**aus; **M**antel – **H**antel

- **Morphem** = kleinste bedeutungstragende Einheiten einer Sprache

  Be-**haus**-ung, **haus**-en, **Haus**-ierer

- **Wort** = bedeutungstragende Einheiten der Sprache, bestehend aus einzelnen Morphemen oder einer Kombination mehrerer Morpheme; abstrakt lexikalisch:

  Haus [Substantiv; *Gebäude*]

  Maus [Substantiv; 1. *Tier*, 2. *PC-Bediengerät*]

  hausen [Verb, *umgspr. für wohnen*]

- **Wortform** = Erscheinungsformen von Wörtern in der Sprache; Zuordnung zur lexikalischen Einheit, z.B.*:*

  Haus, Häuser, Hauses, Häusern etc.

  hausen, hausend

## Bestandteile der Sprache II

- **Wörter** können im Satz ausgetauscht werden und Satzglieder bilden:

  Der Mond ist aus **grünem** Käse.

  Der Mond ist aus **gelbem** Käse.

- **Satzteil, Syntagma** = bedeutungstragende, selbstständige Teile eines Satzes

  Hans schläft. (Subjekt und Prädikat)

  Hans schläft stundenlang in meiner Vorlesung. (Subjekt, Prädikat, Objekt)

- **Satz** = Wortfolge mit mindestens einem Objekt (Subjekt) und einem Prädikat

  Studenten lieben lange Vorlesungen.

  Studenten, die morgens unausgeschlafen sind, weil sie nachts zu lange gearbeitet haben, lieben es, in langen Vorlesungen, die von ihren Professoren spannend und abwechslungsreich dargeboten werden, stundenlang aufmerksam zuzuhören.

## Morphologie – Wörter und ihre Bestandteile

### 3 Klassen von Wörtern

- **einfache Wörter** (Simplizia)

  **Uhr** (Kernmorphem)

  **Uhr** - en (Kernmorphem und Flexionsmorphem)

- **Ableitungen** (Derivationen)

  Ver - **bind** - ung – en
  (KM, ggf. FM und zus. Wortbildungsmorphem(e))

- **Komposita** (mind. 2 KM, ggf. FM, DM und ggf. Fugenelement)

  Uhr - en - ver – gleich - s - test

**Wortbildung**

erfolgt z.B. durch

➢ Hinzufügung von **Präfixen** zum Wortstamm

ver - walt - en
P K F

un - ver - schämt
P P K

➢ Hinzufügung von **Suffixen** zum Wortstamm

Ver - walt - ung
P K S

Ver - un - rein - ig - ung
P P K S S

## Voraussetzung

Der Einsatz eines **regelbasierten Verfahrens** ist nur dann sinnvoll, wenn die Quellsprache über eine im hohen Maße **regelhafte Wortbildung** verfügt, d.h.

- die Zahl der benötigten Regeln nicht zu hoch ist,
- die Zahl der zu erfassenden Ausnahmefälle nicht zu hoch ist.

Beide Bedingungen sind z.B. für das Englische erfüllt.

## Arbeitsweise von Stemmern

1. Entwicklung eines **Sets von Regeln**, mit dem unterschiedliche Fälle von Flexionsendungen unterschieden werden könnnen.
2. Festlegen von **Manipulationen**, die aus **Wortformen** unter Verwendung von (1) **Grundformen** oder **Stämme** generieren.
3. Festlegen einer **Ausnahmeliste**, in die alle nicht regelhaften Fälle eingetragen werden.

## Stemming-Verfahren

Der Stemmer arbeitet mit der folgenden **Abarbeitungsreihenfolge**

1. Versuch einer Identifizierung über Ausnahmeliste
2. Anwendung des Regelwerks,

d. h. für alle Ausnahmen wird das Regelwerk **nicht** aktiviert.

## Ziele

Generierung von **grammatikalischen Grundformen** als Indextermen; Flexions-endungen werden entfernt, die Wortklasse bleibt erhalten (Lexikoneintrag):

*retrieval, retrieve*

Generierung von **Wortstämmen** als Indextermen; Wortbildungsbestandteile (Derivate) werden entfernt, die Wortklasse geht verloren:

*retriev*

Wortstämme und Grundformen können in manchen Fällen auch identisch sein:

*sea*

| | | | |
|---|---|---|---|
| 1. | **IES** | ⇨ | **Y** |
| 2. | **ES** | ⇨ | **_** [wenn *O / CH / SH / SS / ZZ / X vorausgehen] |
| 3. | **S** | ⇨ | **_** [wenn * / E / %Y / %O / OA / EA vorausgehen] |
| 4. | **IES'** | ⇨ | **Y** |
| | **ES'** | ⇨ | **_** |
| | **S'** | ⇨ | **_** |
| 5. | **'S** | ⇨ | **_** |
| | **'** | ⇨ | **_** |
| 6. | **ING** | ⇨ | **_** [wenn ** / % / X vorausgehen] |
| | **ING** | ⇨ | **E** [wenn %* vorausgehen] |
| 7. | **IED** | ⇨ | **Y** |
| 8. | **ED** | ⇨ | **_** [wenn ** / % / X vorausgehen] |
| | **ED** | ⇨ | **E** [wenn %* vorausgehen] |

| | |
|---|---|
| **%** | **= alle Vokale und Y** |
| ***** | **= alle Konsonanten** |
| **_** | **= Tilgung** |
| **/** | **= Oder** |

Der vollständige Kuhlen-Algorithmus erreicht eine Fehlerquote < 3%!

## Voraussetzung

Falls für ein **regelbasiertes Verfahren**

- die Zahl der benötigten Regeln zu hoch wäre **und**
- die Zahl der zu erfassenden Ausnahmefälle zu hoch wäre,

besteht die Alternative in einem **wörterbuchbasierten Verfahren**.
Dies ist typischerweise z.B. für das Deutsche so.

## Arbeitsweise eines wörterbuchbasierten Verfahrens zur Grundformreduktion

1. Aufbau eines **Wörterbuchs** als **Positivliste**, in dem entweder alle Wörter einer Sprache als Grundform oder als Vollform aufgenommen sind.
2. Festlegen einer **Identifizierungsstrategie**, um Wörter in Texten (Wortformen) erkennen und in Grundform bringen zu können.
3. Festlegen eines Verfahrens zur Identifizierung und Zerlegung von **Komposita**.

# Die Wortarten des Deutschen

der
Artikel, bestimmter

Oh
Interjektion

dunkel
Adjektiv

sollen
Modalverb

dritter
Ordinalzahl (Numeral)

lernen
Verb

die (welche)
Relativpronomen

zwei
Kardinalzahl (Numeral)

Heuschrecke
Substantiv/Nomen

vorhin
Adverb

hinter
Präposition

sie
Personalpronomen

haben
Hilfsverb

eine
Artikel, unbestimmter

dieser
Demonstrativpronomen

mein
Possessivpronomen

weil
Konjunktion, subordinierend

und
Konjunktion, koordinierend

## Wörter

**Substantiv/Nomen**

Heuschrecke, Computer, Langeweile, Werner

**Artikel**

bestimmt: der, die, das

unbestimmt: ein, eine, ein

**Pronomen**

*Personalpronomen*

er, sie, es

*Demonstrativpronomen*

dieser, diese, dieses

*Possessivpronomen*

mein, dein, sein

*Relativpronomen*

der, die, das

**Numeral**

eins, zwei, drei (*Kardinalzahlen*)

erster, zweiter, dritter (*Ordinalzahlen*)

**Adjektiv**

groß, lang, dunkel

**Verb**

lernen, arbeiten

haben, werden, sein (*Hilfsverben*)

können, sollen, müssen, dürfen, mögen, wollen (*Modalverben*)

**Adverb**

heute, vorhin, rechts, ungefähr, hoch

**Präposition**

an, auf, hinter, vor

**Konjunktion**

und, oder (*koordinierend*)

weil, nachdem (*subordinierend*)

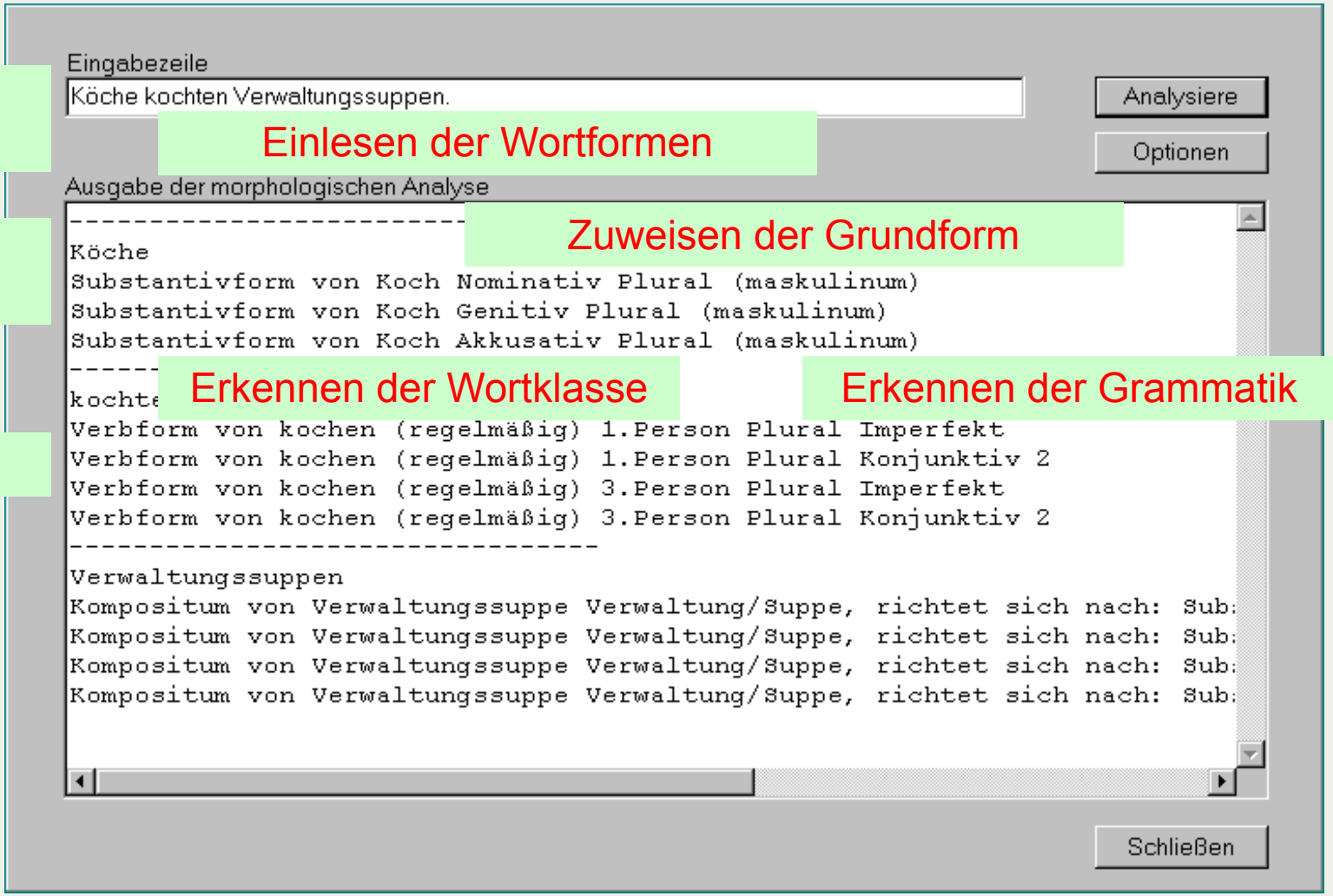
**Interjektion**

oh, au, ach

unregelmäßiger Plural

Verbform Vergangenheit

Kompositum



einfache txt-Datei

```
lahmheit=lahmheit #s
lahmlegend=lahmlegend #a lahmlegen #v
lahmt=lahmen #v
lahmzulegen=lahmlegen #v
lahn=lahn #s
lahne=lahne #s
lahnen=lahnen #v
lahnend=lahnend #a lahnen #v
lahnung=lahnung #s
lahr=lahr #e
laib=laib #s
laibach=laibach #s
laibchen=laibchen #s
laibung=laibung #s
laich=laich #s
laichen=laichen #v
laichend=laichend #a laichen #v
laichingen=laichingen #s
laie=laie #s
laiendarsteller=laiendarsteller #s
laienhaft=laienhaft #a
laientreffen=laientreffen #s
laientum=laientum #s
```

Wortform

Grundform

Wortklasse

„Laien“

=

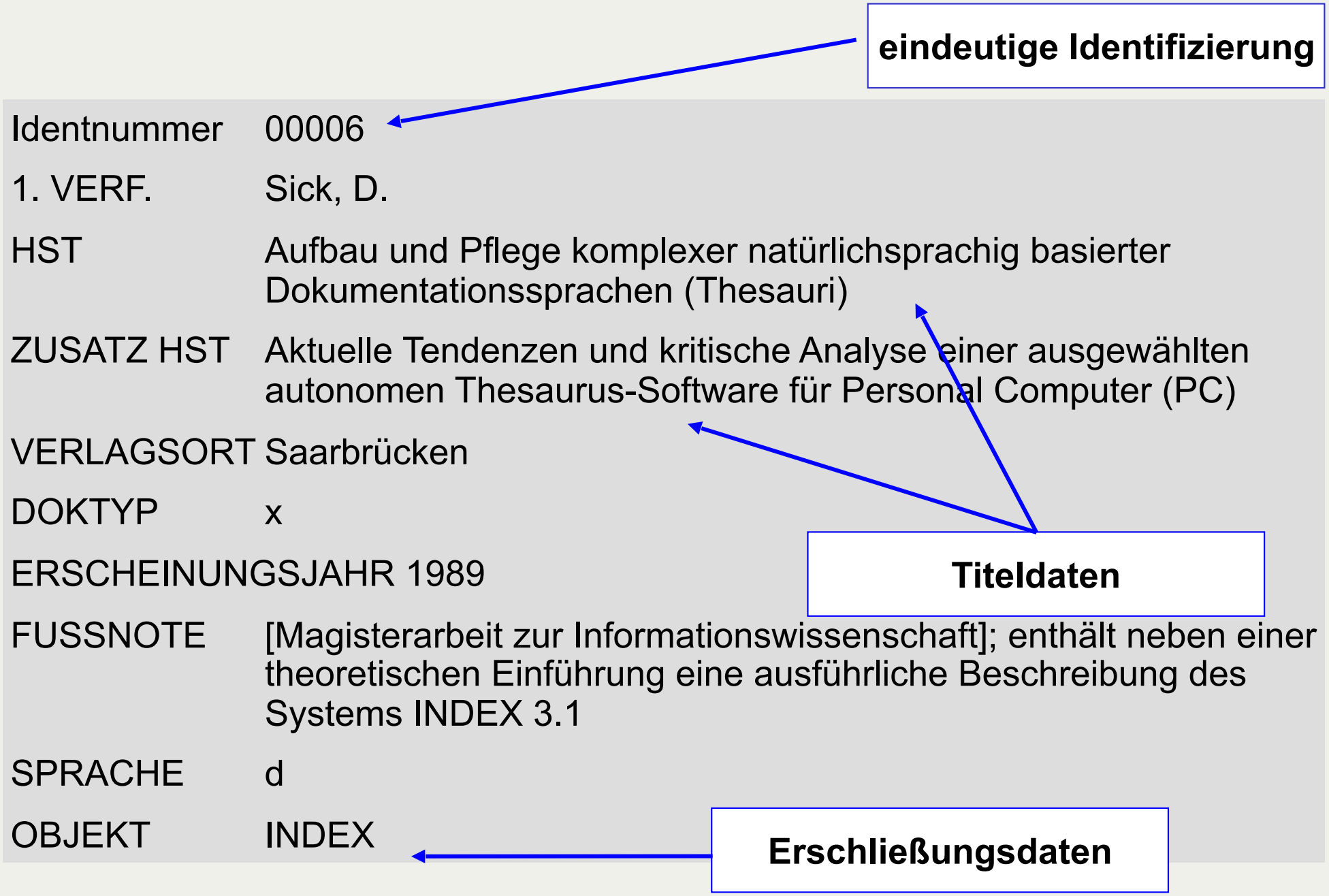
Laie

+

n

```
lahmheit=lahmheit #s
lahmlegend=lahmlegend #a lahmlegen #v
lahmt=lahmen #v
lahmzulegen=lahmlegen #v
lahn=lahn #s
lahne=lahne #s
lahnen=lahnen #v
lahnend=lahnend #a lahnen #v
lahnung=lahnung #s
lahr=lahr #e
laib=laib #s
laibach=laibach #s
laibchen=laibchen #s
laibung=laibung #s
laich=laich #s
laichen=laichen #v
laichend=laichend #a laichen #v
laichingen=laichingen #s
laie=laie #s
laiendarsteller=laiendarsteller #s
laienhaft=laienhaft #a
laientreffen=laientreffen #s
laientum=laientum #s
```

```
suffix:
      #    Suffixliste, Stand: 30-06-2005
      #    Suffixklasse:    s = Substantiv, a = Adjektiv, v = Verb, e = Eigenwort, f = Fugung
      #    Suffixe je Klasse:    "<suffix>['/'<ersetzung>][ <suffix>['/'<ersetzung>]]"
      - [s, "e en er ern es n s se sen ses"]
      - [a, "este ste ster sten stes ester estes esten e em en er ere eren erer eres es"]
      - [v, "e/en en/en est/en et/en st/en t/en te/en ten/en eten/en ete/en etest/en s"]
      - [e, "s"]
      - [f, "s n e en es er"]
```

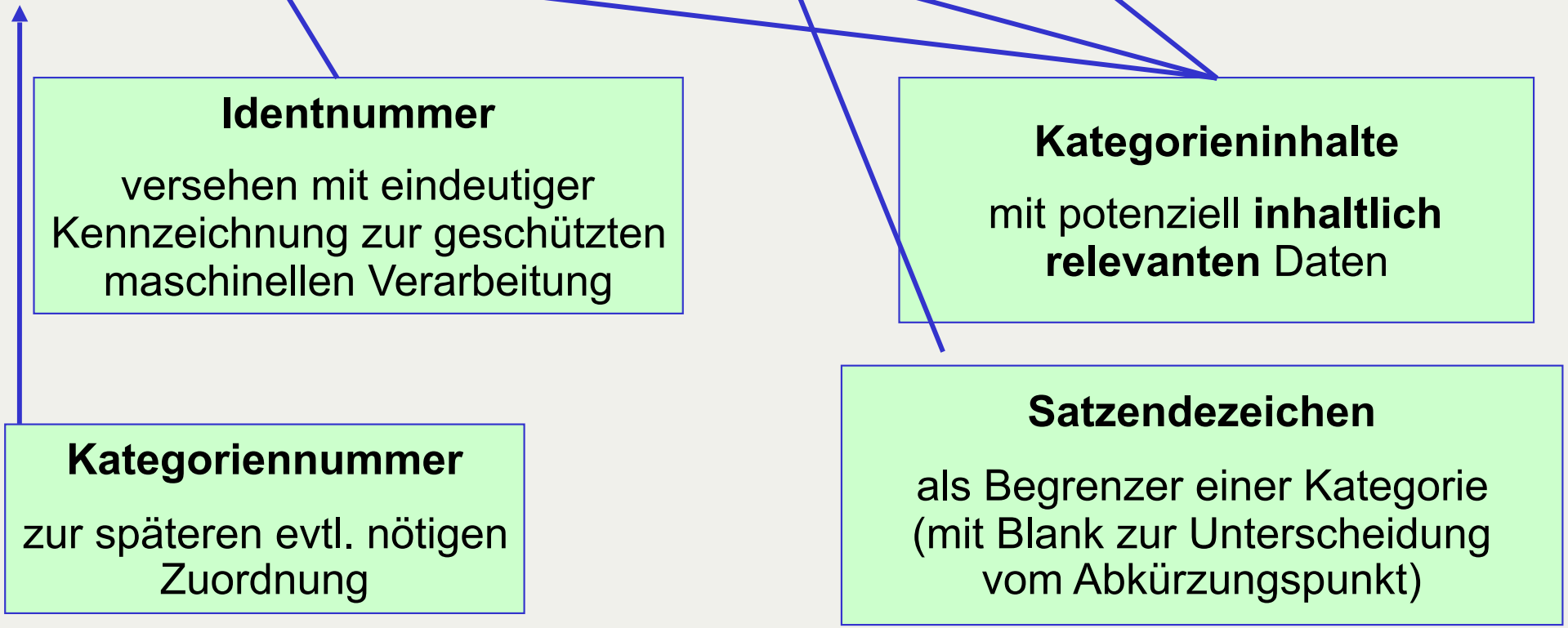

eindeutige Identifizierung
Identnummer 00006
1. VERF. Sick, D.
HST Aufbau und Pflege komplexer natürlichsprachig basierter Dokumentationssprachen (Thesauri)
ZUSATZ HST Aktuelle Tendenzen und kritische Analyse einer ausgewählten autonomen Thesaurus-Software für Personal Computer (PC)
VERLAGSORT Saarbrücken
DOKTYP x
ERSCHEINUNGSJAHR 1989
FUSSNOTE [Magisterarbeit zur Informationswissenschaft]; enthält neben einer theoretischen Einführung eine ausführliche Beschreibung des Systems INDEX 3.1
SPRACHE d
OBJEKT INDEX
Titeldaten
Erschließungsdaten

[00006 .]

020: Aufbau und Pflege komplexer natürlichsprachig basierter Dokumentationssprachen (Thesauri) .

025: Aktuelle Tendenzen und kritische Analyse einer ausgewählten autonomen Thesaurus-Software für Personal Computer (PC) .

100: INDEX .



```
lex:)  :[/OTHR:
lex:)  :00006./NUMS:
lex:)  :]/OTHR:
lex:)  :020/NUMS:
lex:)  ::/PUNC:
lex:)  <Aufbau = [(aufbau/s)]>
lex:)  <und = [(und/w)]>
lex:)  <Pflege = [(pflege/s)]>
lex:)  <komplexer = [(komplex/s), (komplex/a)]>
lex:)  <natürlichsprachig|?>
lex:)  <basierter = [(basiert/a)]>
lex:)  <Dokumentationssprachen|KOM = [(dokumentationssprache/k), (dokumentation/s+), (sprache/s+)]>
lex:)  :(/OTHR:
lex:)  <Thesauri|?>
lex:)  :025/NUMS:
lex:)  <Aktuelle = [(aktuell/a)]>
lex:)  <Tendenzen = [(tendenz/s)]>
lex:)  <und = [(und/w)]>
lex:)  <kritische = [(kritisch/a)]>
lex:)  <Analyse = [(analyse/s), (analytik/y)]>
lex:)  <einer = [(einer/s), (ein/w), (einer/w)]>
lex:)  <ausgewählten = [(ausgewählt/a)]>
lex:)  <autonomen = [(autonom/a)]>
lex:)  <Thesaurus-Software|KOM = [(thesaurus-software/k), (software/s+), (thesaurus/s+)]>
lex:)  <für = [(für/w)]>
lex:)  <Personal = [(personal/s), (personal/a)]>
lex:)  <Computer = [(computer/s)]>
```

**Identnummer**

**Wortklasse**

**Grundformen**

00006*aktuell | analyse | aufbau | ausgewählt | autonom | basiert | computer | dokumentationssprache | einer | komplex | kritisch | personal | pflege | tendenz | thesaurus-software

| | |
|---|---|
| Identnummer | 00006 |
| 1. VERF. | Sick, D. |
| HST | Aufbau und Pflege komplexer natürlichsprachig basierter Dokumentationssprachen (Thesauri) |
| ZUSATZ HST | Aktuelle Tendenzen und kritische Analyse einer ausgewählten autonomen Thesaurus-Software für Personal Computer (PC) |
| VERLAGSORT | Saarbrücken |
| DOKTYP | x |
| ERSCHEINUNGSJAHR | 1989 |
| FUSSNOTE | [Magisterarbeit zur Information theoretischen Einführung eine ausführliche Beschreibung des Systems INDEX 3.1 |
| SPRACHE | d |
| OBJEKT | INDEX |
| Indexate | 00006*aktuell \| analyse \| aufbau \| ausgewählt \| autonom \| basiert \| computer \| dokumentationssprache \| einer \| komplex \| kritisch \| personal \| pflege \| tendenz \| thesaurus-software |

**Speicherformat**

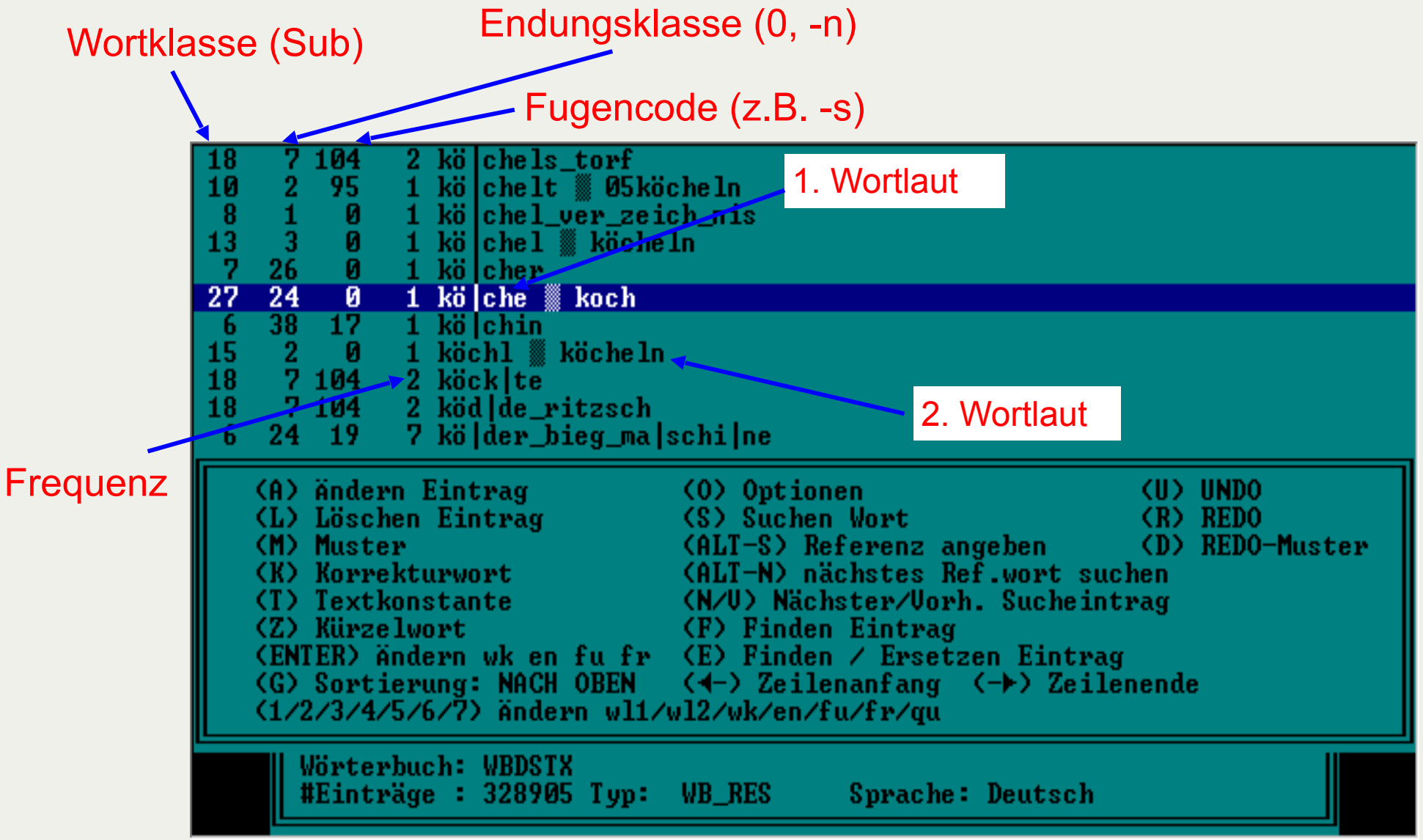
(Komma delimited)

**Import**

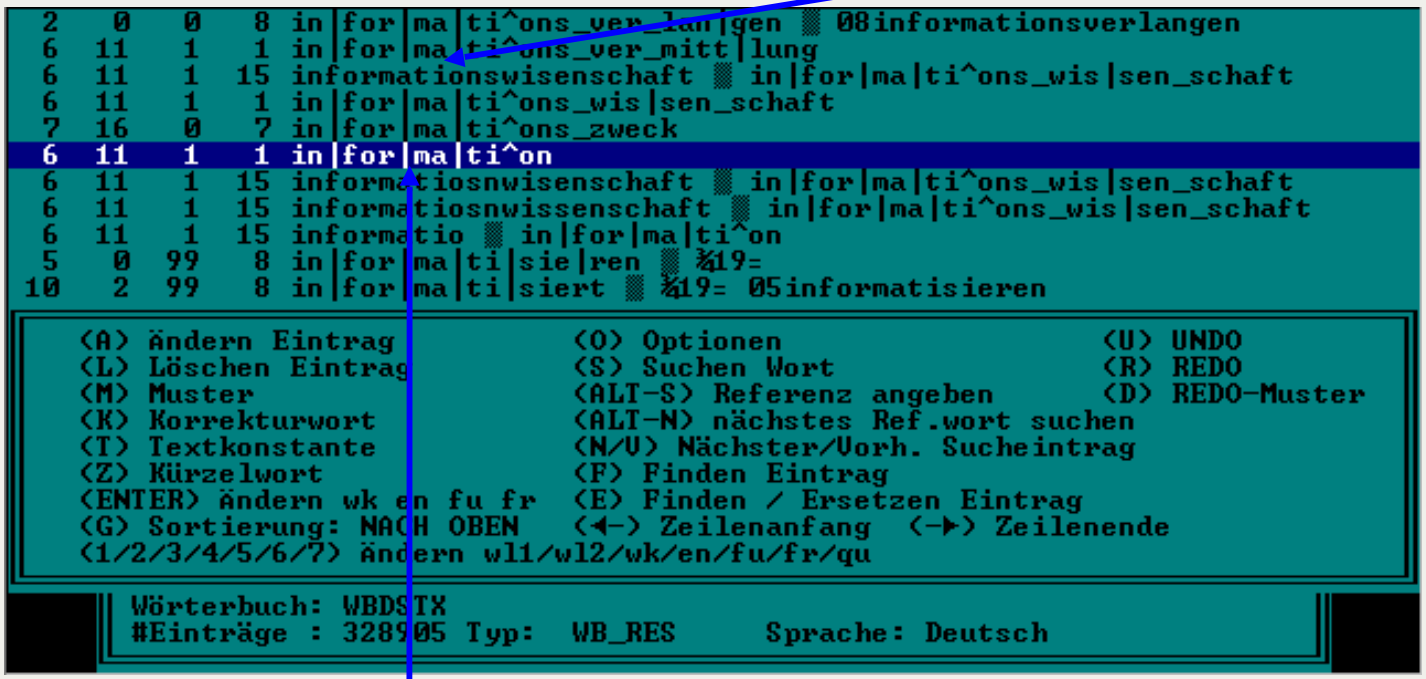
in zusätzliche Kategorie

(kann für den Indexaufbau genutzt werden)

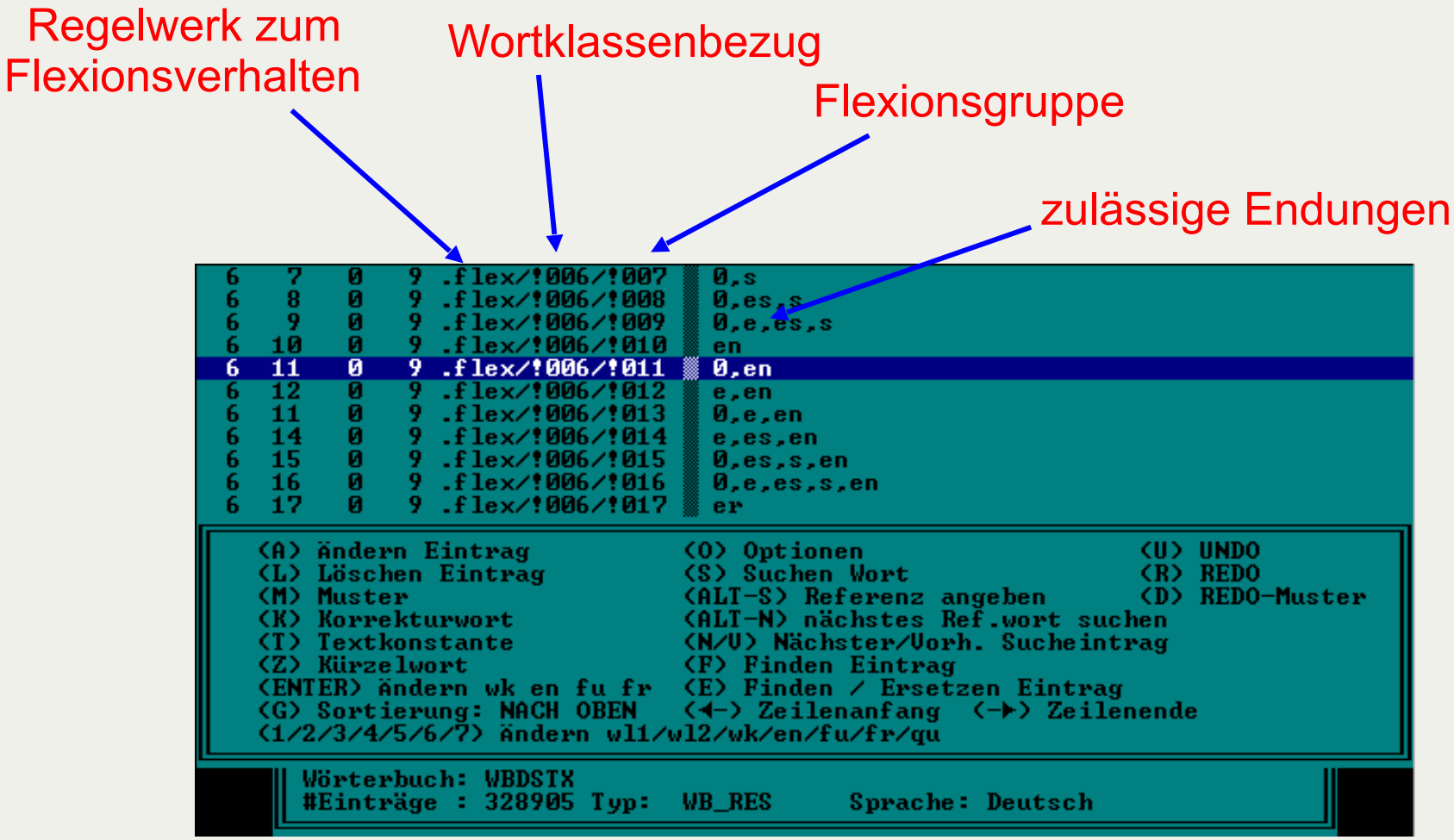

```
Abfall::Abfall+SUB;MAS;SG;AKK
Abfall::Abfall+SUB;MAS;SG;DAT
Abfall::Abfall+SUB;MAS;SG;NOM
Abfalles::Abfall+SUB;MAS;SG;GEN
Abfalls::Abfall+SUB;MAS;SG;GEN
Abfaltung::Abfaltung+SUB;FEM;SG;AKK
Abfaltung::Abfaltung+SUB;FEM;SG;DAT
Abfaltung::Abfaltung+SUB;FEM;SG;GEN
Abfaltung::Abfaltung+SUB;FEM;SG;NOM
Abfaltungen::Abfaltung+SUB;FEM;PL;AKK
Abfaltungen::Abfaltung+SUB;FEM;PL;DAT
Abfaltungen::Abfaltung+SUB;FEM;PL;GEN
Abfaltungen::Abfaltung+SUB;FEM;PL;NOM
Abfassung::Abfassung+SUB;FEM;SG;AKK
Abfassung::Abfassung+SUB;FEM;SG;DAT
Abfassung::Abfassung+SUB;FEM;SG;GEN
Abfassung::Abfassung+SUB;FEM;SG;NOM
Abfassungen::Abfassung+SUB;FEM;PL;AKK
Abfassungen::Abfassung+SUB;FEM;PL;DAT
Abfassungen::Abfassung+SUB;FEM;PL;GEN
Abfassungen::Abfassung+SUB;FEM;PL;NOM
```


Wortklasse (Sub)
Endungsklasse (0, -n)
Fugencode (z.B. -s)
1. Wortlaut
2. Wortlaut
Frequenz
18 7 104 2 kö|chels_torf
10 2 95 1 kö|chelt ▒ 05köcheln
8 1 0 1 kö|chel_ver_zeich_nis
13 3 0 1 kö|chel ▒ köcheln
7 26 0 1 kö|cher
27 24 0 1 kö|che ▒ koch
6 38 17 1 kö|chin
15 2 0 1 köchl ▒ köcheln
18 7 104 2 köck|te
18 7 104 2 köd|de_ritzsch
6 24 19 7 kö|der_bieg_ma|schi|ne
(A) Ändern Eintrag
(L) Löschen Eintrag
(M) Muster
(K) Korrekturwort
(T) Textkonstante
(Z) Kürzelwort
(ENTER) ändern wk en fu fr
(G) Sortierung: NACH OBEN
(1/2/3/4/5/6/7) ändern wl1/wl2/wk/en/fu/fr/qu
(O) Optionen
(S) Suchen Wort
(ALT-S) Referenz angeben
(ALT-N) nächstes Ref.wort suchen
(N/V) Nächster/Vorh. Sucheintrag
(F) Finden Eintrag
(E) Finden / Ersetzen Eintrag
(←) Zeilenanfang (→) Zeilenende
(U) UNDO
(R) REDO
(D) REDO-Muster
Wörterbuch: WBDSTX
#Einträge : 328905 Typ: WB_RES Sprache: Deutsch

Beachte "Longest-Matching-Sortierung"




Eingabestring "Informationen" führt zu Lexikoneintrag "Information"




Grundform "Information" + Wortklasse Substantiv + Endung "en"
= "Informationen" (Wortform)

Zerlegungsversuch über "Information"

Informationswirtschaft

nicht im Lexikon!

informatik=informatik #s
informatiker=informatiker #s
informatiksystem=informatiksystem #s
information=information #s
informationsministerium=informationsministerium #s
informationsoffizier=informationsoffizier #s
informationsverarbeitend=informationsverarbeitend #a
informativ=informativ #a
informatorisch=informatorisch #a
informell=informell #a
informieren=informieren #v
informierend=informierend #a informieren #v
informiert=informieren #v informiert #a
infostand=infostand #s

Fugencode erlaubt
Fugen-s

Suchstring "Information-s" ist identifiziert

```
suffix:
      #    Suffixliste, Stand: 30-06-2005
      #    Suffixklasse:   s = Substantiv, a = Adjektiv, v = Verb, e = Eigenwort, f = Fugung
      #    Suffixe je Klasse:   "<suffix>['/'<ersetzung>][ <suffix>['/'<ersetzung>]]"
      - [s, "e en er ern es n s se sen ses"]
      - [a, "este ste ster sten stes ester estes esten e em en er ere eren erer eres es"]
      - [v, "e/en en/en est/en et/en st/en t/en te/en ten/en eten/en ete/en etest/en s"]
      - [e, "s"]
      - [f, "s n e en es er"]
```

Fortsetzung der Identifizierung mit "Wirtschaft" und Beenden der Kompositumanalyse

**Aber Achtung!**

Warum nicht Zerlegung von "Wirtschaft" in "Wirt" und "Schaft" ?

wirt=wirt #s
wirten=wirt #s wirten #v
wirtend=wirtend #a
wirtlich=wirtlich #a
wirtschaft=wirtschaft #s
wirtschaften=wirtschaft #s wirtschaften #v
wirtschaftend=wirtschaftend #a wirtschaften #v
wirtschafter=wirtschafter #s

lex:) <informationswissenschaft|KOM = [(informationswissenschaft/k), (information/s+), (wissenschaft/s+)]>
lex:) <informationswirtschaft|KOM = [(informationswirtschaft/k), (information/s+), (wirtschaft/s+)]>
lex:) <wissenschaft = [(wissenschaft/s)]>
lex:) <wirtschaft = [(wirtschaft/s)]>

Was ist mit sprachlichen Problemfällen, z.B. mit mehrdeutigen Komposita?

"Baumangel"

Zerlegung in "Baum" und "Angel" oder in "Bau" und "Mangel"? Oder beides?

# Wortklassenkennung (+) von Zerlegungsergebnissen

lex:) <Informationsaufkommens|KOM = [(informationsaufkommen/k), (aufkommen/s+), (information/s+), (aufkommen/v+)]>

lex:) <Häufigkeitsverteilung|KOM = [(häufigkeitsverteilung/k), (häufigkeit/s+), (verteilung/s+)]>

lex:) <Erzeugnisbeschreibungen|KOM = [(erzeugnisbeschreibung/k), (beschreibung/s+), (erzeugnis/s+)]>

lex:) <Vektorraum-Modell|KOM = [(vektorraum-modell/k), (modell/s+), (raum/s+), (vektor/s+)]>

lex:) <Information-Retrieval-Systemen|KOM = [(information-retrieval-system/k), (information/s+), (system/s+), (retrieval/x+)]>

lex:) <Dokumentenbeständen|KOM = [(dokumentenbestand/k), (bestand/s+), (bestände/s+), (dokument/s+)]>

lex:) <Termunabhängigkeitsannahme|KOM = [(termunabhängigkeitsannahme/k), (annahme/s+), (term/s+), (unabhängigkeit/s+)]>

lex:) <Sacherschließung|KOM = [(sacherschließung/k), (sacher/e+), (schließung/s+)]>

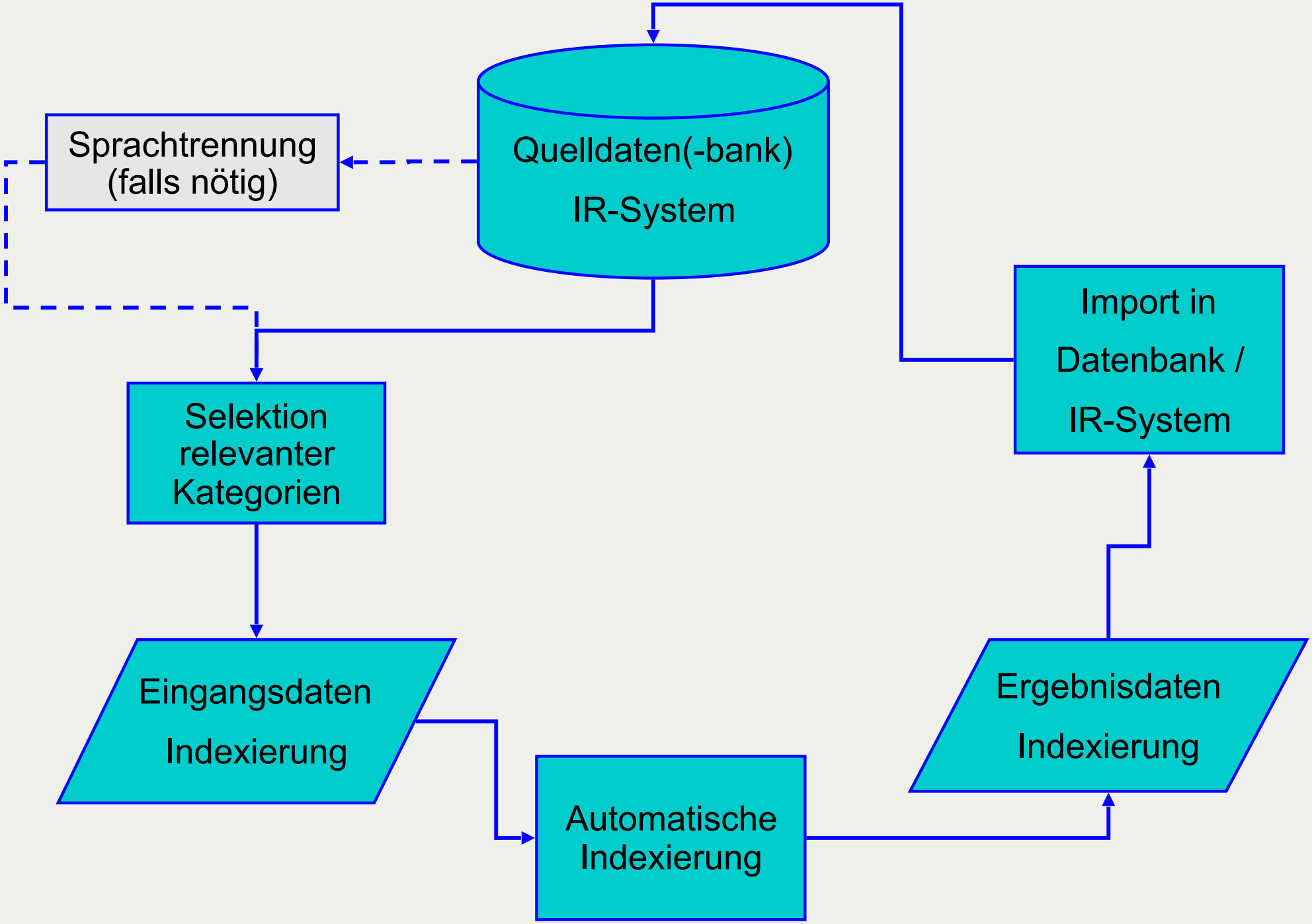
lex:) <Inhaltserschließung|KOM = [(inhaltserschließung/k), (erschließung/s+), (inhalt/s+)]>

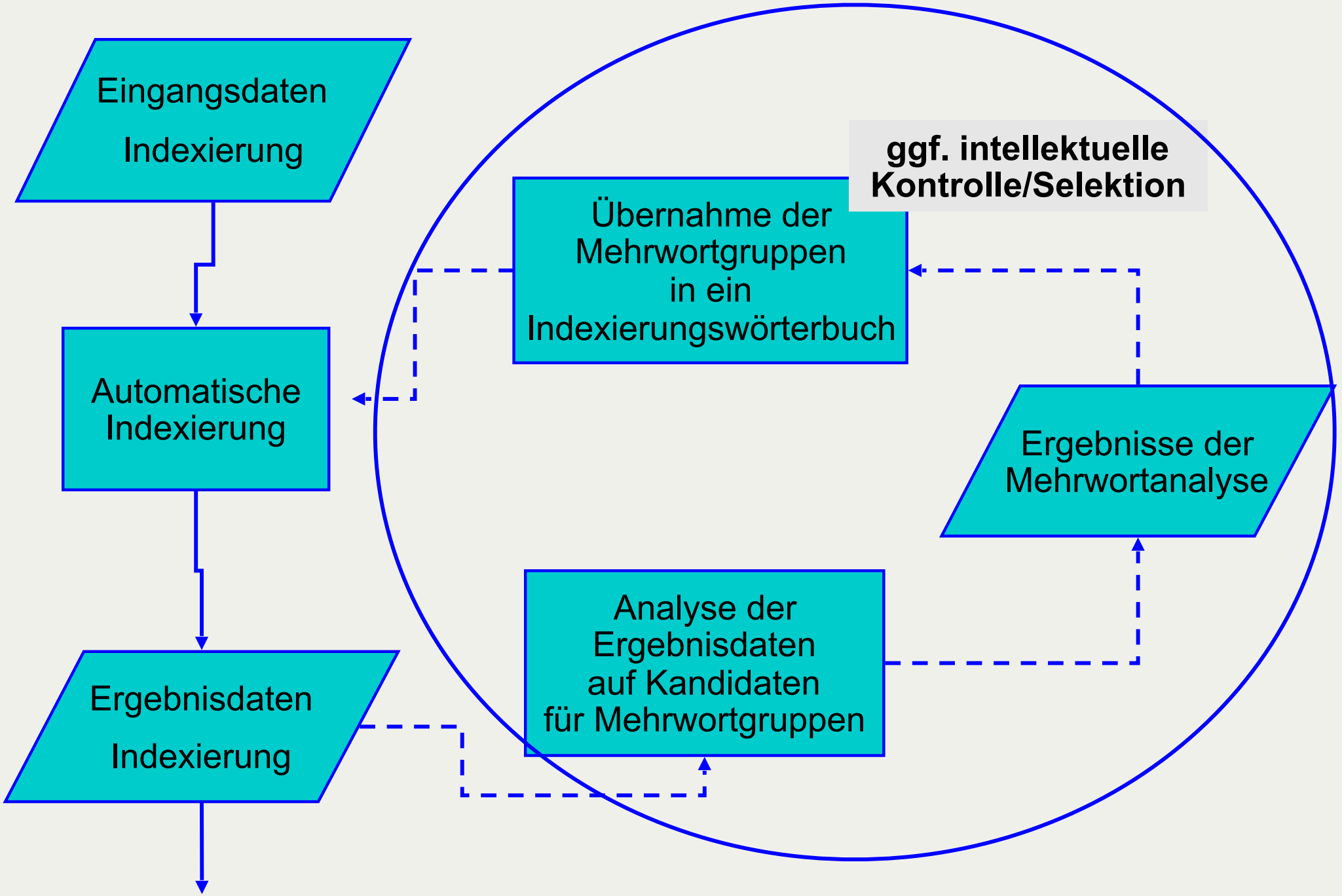
Bindestrichwörter wie Komposita

dreigliedriges Kompositum

„sach“ nicht im Wörterbuch; Zerlegungskonflikt!

beide Bestandteile im Wörterbuch longest matching!


Quelldaten(-bank)
IR-System
Sprachtrennung
(falls nötig)
Selektion
relevanter
Kategorien
Eingangsdaten
Indexierung
Automatische
Indexierung
Ergebnisdaten
Indexierung
Import in
Datenbank /
IR-System


Eingangsdaten
Indexierung
Automatische
Indexierung
Ergebnisdaten
Indexierung
ggf. intellektuelle
Kontrolle/Selektion
Übernahme der
Mehrwortgruppen
in ein
Indexierungswörterbuch
Ergebnisse der
Mehrwortanalyse
Analyse der
Ergebnisdaten
auf Kandidaten
für Mehrwortgruppen

# Lingo - sequencer

automatisch akquisition -- akquisition, automatisch
automatisch bestimmung -- bestimmung, automatisch
automatisch dokumenterschließung -- dokumenterschließung, automatisch
automatisch dokumentklassifikation -- dokumentklassifikation, automatisch
automatisch erschließung -- erschließung, automatisch
automatisch gruppierung -- gruppierung, automatisch
automatisch indexieren -- indexieren, automatisch
automatisch indexierung -- indexierung, automatisch
automatisch indexierungssystem -- indexierungssystem, automatisch
automatisch inhaltlich erschließung -- erschließung, automatisch inhaltlich
automatisch inhaltserschließung -- inhaltserschließung, automatisch
automatisch klassifikation -- klassifikation, automatisch
automatisch maschine -- maschine, automatisch
automatisch methode -- methode, automatisch
automatisch recherche -- recherche, automatisch
automatisch selektion -- selektion, automatisch
automatisch semantisch klassifikation -- klassifikation, automatisch semantisch
automatisch thematisch textklassifikation -- textklassifikation, automatisch thematisch
automatisch verfahren -- verfahren, automatisch
automatisch vollindexierung -- vollindexierung, automatisch
automatisch wortformenreduktion -- wortformenreduktion, automatisch
automatisch übersetzungssystem -- übersetzungssystem, automatisch

**Verbindungen**

von

**Adjektiv** und **Substantiv**

**Verbindungen**

von

**Adjektiv**, **Adjektiv** und **Substantiv**

# Lingo - synonymer

Thesauruseintrag

```
Abfallbeseitigung
    Q M  SYS 31.2
    BF Abfallentsorgung
    BF Hausmüllentsorgung
    BF Müllbeseitigung
    OB ^Entsorgung
```

**erzeugt folgende Einträge in einem Relationenwörterbuch**

- Abfallentsorgung ⇨ Abfallbeseitigung
- Hausmüllentsorgung ⇨ Abfallbeseitigung
- Müllbeseitigung ⇨ Abfallbeseitigung
- Abfallbeseitigung ⇨ Entsorgung
- [Abfallbeseitigung ⇨ Abfallentsorgung]
- [Abfallbeseitigung ⇨ Hausmüllentsorgung]
- [Abfallbeseitigung ⇨ Müllbeseitigung]
- [Entsorgung ⇨ Abfallbeseitigung]
- [Entsorgung ⇨ Abfallentsorgung]
- [Entsorgung ⇨ Hausmüllentsorgung]
- [Entsorgung ⇨ Müllbeseitigung]
- [Abfallentsorgung ⇨ Hausmüllentsorgung]
- [Abfallentsorgung ⇨ Müllbeseitigung]
- [Hausmüllentsorgung ⇨ Abfallentsorgung]
- [Hausmüllentsorgung ⇨ Müllbeseitigung]
- [Müllbeseitigung ⇨ Abfallentsorgung]
- [Müllbeseitigung ⇨ Hausmüllentsorgung]

durch Thesauruseinträge erzeugte Relationierungen (Kennung y)

lex:) <Dokumentationssprachen|KOM = [(dokumentationssprache/k), (indexierungssprache/y), (informationssprache/y), (dokumentation/s+), (sprache/s+)]>
lex:) <Wörterbüchern = [(wörterbuch/s), (wörterbücher/s), (sprachwörterbuch/y), (vokabularium/y)]>
lex:) <Transparenz = [(transparenz/s), (durchsichtigkeit/y), (optische transparenz/y), (translucency/y), (transluzenz/y), (transparency/y)]>
lex:) <Klassifikation = [(klassifikation/s), (klassenbildung/y), (klassifikationssystem/y), (klassifizierung/y)]>
lex:) <Computers = [(computer/s), (digitale datenverarbeitungsanlage/y), (digitale rechenanlage/y), (digitaler computer/y), (digitalrechner/y), (dva/y), (elektronenrechner/y), (elektronische rechenanlage/y), (elektronischer rechenautomat/y), (programmgesteuerter digitaler rechenautomat/y), (rechenanlage/y), (rechenautomat/y), (rechner/y)]>
lex:) <Leistungsbewertung|KOM = [(leistungsbewertung/k), (leistungsanalyse/y), (performance analysis/y), (performance evaluation/y), (performance bewertung/y), (performancebewertung/y), (bewertung/s
lex:) <Information-Re
(dokumentenretrievalsystem/y), (dokumentensuchsystem/y), (dokumentsuchsystem/y), (informationswiedergewinnungssystem/y), (ir-system/y), (retrievalsystem/y), (textretrievalsystem/y), (information/s+), (system/s+), (retrieval/x+)]>
lex:) <Bibliographie = [(bibliographie/s), (bibliografie/y), (bücherverzeichnis/y), (literaturverzeichnis/y)]>

Ergebnis: zusätzliche Sucheinstiege im semantischen Umfeld der Wortform!

# Syntaxanalyse, Parsing

**Das Problem**:

Stichwortextraktionsverfahren sind beschränkt auf das Vokabular in den zu indexierenden Kategorien (z.B. Titel)

**Beispiel**:

"Sacherschließung mit Schlagwörtern"

Stichwortextraktion: *Sacherschließung, Schlagwort*

Deskriptoren (intellektuell): *Inhaltserschließung, Deskriptor, verbale Inhaltserschließung*

**Ziele "mächtigerer" automatischer Verfahren**:

- Analyse des Kontexts:
  "mit" Schlagwörtern = *verbale Inhaltserschließung*
- Analyse des "semantischen Umfelds":
  "Sacherschließung" = *Inhaltserschließung*
  "Schlagwort" = *Indexing*

**Verfahren zur Realisierung**

- Syntaxanalyse (Parsing)
- automatische Indexierung mit Thesaurusrelationen

**Parsing** eines Satzes bedeutet, eine Folge von Ableitungen bzw. Regeln zu finden, die von einem (definierten) Startsymbol zum Satz führen.

Parser bedienen sich dazu sog. **formaler Grammatiken**, d.h. **Regelwerken**, die dem Programm angeben, aus welchen Elementen sich **gültige Sätze** zusammensetzen.

Parser sind damit in der Lage

- zu entscheiden, ob ein gegebener Satz ein gültiger ist,
- die grammatikalische Struktur eines Satzes vollständig zu analysieren.

# Bestandteile einer einfachen formalen Grammatik

**1. Symbole** (non-terminale, präterminale, terminale)

- **S:** Startsymbol
- **NP:** Nominalphrase
- **VP:** Verbalphrase
- **PP:** Präpositionalphrase
- **n:** Nomen
- **v:** Verb
- **p:** Präposition

**2. Lexikon:** {er [n], ihn [n], findet [v], sucht [v], im [p], haus [n]}

**3. Regelwerk:**

- **(G1):** S ⇨ NP VP
- **(G2):** NP ⇨ n
- **(G3):** VP ⇨ v NP PP
- **(G4):** PP ⇨ p NP

Die Grammatik erzeugt z.B. folgende Sätze:

*Er sucht ihn im Haus.*

*Ihn findet er im Haus.*

**aber auch:**

*Ihn sucht ihn im ihn.*

*Er findet er im er.*

*Haus findet er im ihn.*

# Umsetzung der Strategie Top-down für den Satz:

*Er sucht ihn im Haus.*

**Strategie für Top-down-Parsing**

**(P1)** Beginne mit dem Startsymbol.

**(P2)** Ersetze das erste nicht-terminale Symbol durch die rechte Seite einer Regel, deren linke Seite mit diesem Symbol identisch ist.

**(P3)** Entferne führende terminale Symbole.

**(P4)** Wenn es noch nicht-terminale Symbole gibt, dann gehe zu (P2).

(1) **S**
beginne mit dem Startsymbol **S** gemäß **(P1)**

(2) **NP1 VP**
ersetze **S** durch **NP VP** gemäß **(P2)** und **(G1)**

(3) **n VP**
wende **(G2)** an, d.h. ersetze **NP** durch **n**

(4) **VP**
entferne das führende terminale Symbol **n** gemäß **(P3)**

(1) **v NP2 PP**
wende **(G2)** an, d.h. ersetze **VP** durch **v NP PP**

(2) **NP2 PP**
entferne das führende terminale Symbol **v** gemäß **(P3)**

(3) **n PP**
wende **(G2)** an, d.h. ersetze **NP** durch **n**

(4) **PP**
wende **(P4)** an, d.h. gehe zu **(P2)**

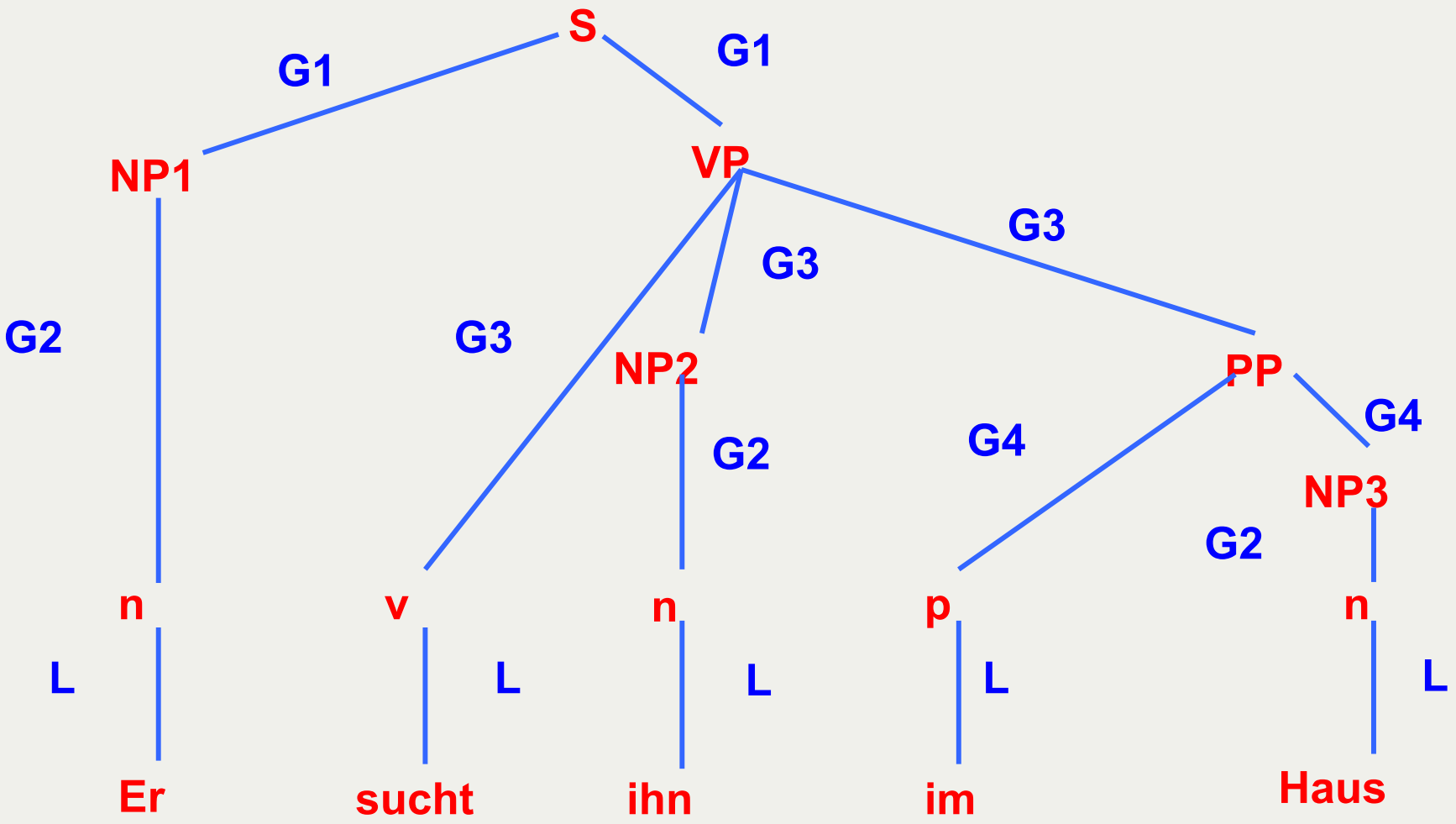
(5) **p NP3**
wende **(G4)** an, d.h. ersetze **PP** durch **p NP**

(6) **NP3**
entferne das führende terminale Symbol **p** gemäß **(P3)**

(7) **n**
wende **(G2)** an, d.h. ersetze **NP** durch **n**

(8) Beende das Parsing gemäß **(P4)**

**Satz:** Er sucht ihn im Haus



# Übung

Gegeben sei folgende Grammatik

**(G1)** S ⇨NPVP

**(G2)** NP ⇨ n

**(G3)** VP ⇨ v NP

folgendes Lexikon

**Lexikon:** {ich [n], esse [v], Käse [n]}

und folgender Satz

*Ich esse Käse.*

Geben Sie den Parsingverlauf für Top-down-Parsing an.

Geben Sie den Parsingverlauf für folgenden Satz an:

*Ich esse grünen Käse.*

## Indexierung eines Beispieldokuments mit Lingo

```
<00001 .>

Gesellschaft: Strahlenrisiko wird drastisch unterschätzt =

    Bremen (dpa) - Eine drastische Fehleinschätzung des
Strahlenrisikos hat die Gesellschaft für Strahlenschutz der
Wirtschaft, der Politik und einer `industriefreundlichen Wissenschaft"
vorgeworfen. Dies habe dazu beigetragen, dass es in Deutschland heute
mehr als 30 000 anerkannte Fälle von Berufskrankheiten gebe, die durch
Arbeiten im Bereich der Atomindustrie unter mangelhaften
Schutzbestimmungen hervor gerufen worden seien, kritisierte der
Präsident der Gesellschaft, Sebastian Pflugbeil (Berlin).

    Die Strahlenschutzverordnung des Bundes habe seit Jahrzehnten ein
unterschätztes Risiko zur Grundlage, sagte Pflugbeil am Donnerstag in
Bremen. Dort beginnt am Freitag der zweitägige internationale Kongress
`Strahlenschutz nach der Jahrtausendwende". Er forderte eine deutliche
Senkung des Grenzwertes für beruflich von Strahlen betroffene
Personen.


dpa/lni sm yyni ba ub

081351 Jun 00
```

```
lex:)  *FILE('txt/dpa.txt')
lex:)  :</OTHR:
lex:)  :00001/NUMS:
lex:)  :./PUNC:
lex:)  :>/OTHR:
lex:)  <Gesellschaft = [(gesellschaft/s)]>
lex:)  ::/PUNC:
lex:)  <Strahlenrisiko|KOM = [(strahlenrisiko/k), (radiation hazard/y), (strahlengefährdung/y),
         (strahlungsgefährdung/y), (strahlungsrisiko/y), (risiko/s+), (strahl/s+), (strahlen/v+)]>
lex:)  <wird = [(werden/v)]>
lex:)  <drastisch = [(drastisch/a)]>
lex:)  <unterschätzt = [(unterschätzen/v)]>
lex:)  :=/OTHR:
lex:)  <Bremen = [(bremen/e)]>
lex:)  :(/OTHR:
lex:)  <dpa|?>
lex:)  :)/OTHR:
lex:)  <-|?>
lex:)  <Eine = [(ein/w)]>
lex:)  <drastisch fehleinschätzung -- fehleinschätzung, drastisch|SEQ = [(drastisch fehleinschätzung --
         fehleinschätzung, drastisch/q)]>
lex:)  <drastische = [(drastisch/a)]>
lex:)  <Fehleinschätzung = [(fehleinschätzung/s)]>
lex:)  <des = [(des/t)]>
```

```
lex:)  <des = [(des/t)]>
lex:)  <Strahlenrisikos|KOM = [(strahlenrisiko/k), (radiation hazard/y), (strahlengefährdung/y),
        (strahlungsgefährdung/y), (strahlungsrisiko/y), (risiko/s+), (strahl/s+), (strahlen/v+)]>
lex:)  <hat = [(hat/t)]>
lex:)  <die = [(die/t)]>
lex:)  <Gesellschaft = [(gesellschaft/s)]>
lex:)  <für = [(für/w)]>
lex:)  <Strahlenschutz|KOM = [(strahlenschutz/k), (strahlenschutzvorsorge/y), (schutz/s+), (strahl/s+),
        (strahlen/v+)]>
lex:)  <der = [(der/t)]>
lex:)  <Wirtschaft = [(wirtschaft/s), (wirtschaftsleben/y)]>
lex:)  :,/PUNC:
lex:)  <der = [(der/t)]>
lex:)  <Politik = [(politik/s), (politische entwicklung/y), (politische lage/y), (staatspolitik/y)]>
lex:)  <und = [(und/w)]>
lex:)  <einer = [(einer/s), (ein/w), (einer/w)]>
lex:)  :`/OTHR:
lex:)  <industriefreundlichen|KOM = [(industriefreundlich/k), (freundlich/a+), (industrie/s+)]>
```

```
lex:)  <Wissenschaft = [(wissenschaft/s), (bürgerliche wissenschaft/y), (wissenschaften/y)]>
lex:)  :"/OTHR:
lex:)  <vorgeworfen = [(vorgeworfen/a)]>
lex:)  :./PUNC:
lex:)  <Dies = [(dies/w)]>
lex:)  <habe = [(haben/v)]>
lex:)  <dazu = [(dazu/w)]>
lex:)  <beigetragen = [(beitragen/v), (beigetragen/a)]>
lex:)  :,/PUNC:
lex:)  <dass = [(daß/w)]>
lex:)  <es = [(es/t)]>
lex:)  <in = [(in/t)]>
lex:)  <Deutschland = [(deutschland/e)]>
lex:)  <heute = [(heute/w)]>
lex:)  <mehr = [(mehr/s), (mehr/w)]>
lex:)  <als|?>
lex:)  :30/NUMS:
lex:)  :000/NUMS:
lex:)  <anerkannte = [(anerkannt/a)]>
lex:)  <Fälle = [(fällen/v)]>
lex:)  <von|?>
```

```
lex:)  <Berufskrankheiten|KOM = [(berufskrankheit/k), (arbeitsbedingte krankheit/y), (beruf/s+), (krankheit/
        s+)]>
lex:)  <gebe = [(geben/v)]>
lex:)  :,/PUNC:
lex:)  <die = [(die/t)]>
lex:)  <durch = [(durch/w)]>
lex:)  <Arbeiten = [(arbeit/s), (arbeiten/v), (erwerbsarbeit/y)]>
lex:)  <im|?>
lex:)  <Bereich = [(bereich/s)]>
lex:)  <der = [(der/t)]>
lex:)  <Atomindustrie|KOM = [(atomindustrie/k), (kerntechnische industrie/y), (atom/s+), (industrie/s+)]>
lex:)  <unter = [(unter/w)]>
lex:)  <mangelhaften|KOM = [(mangelhaft/k), (haft/s+), (mangel/s+), (haften/v+)]>
lex:)  <Schutzbestimmungen|KOM = [(schutzbestimmung/k), (bestimmung/s+), (schutz/s+)]>
lex:)  <hervor = [(hervor/w)]>
lex:)  <gerufen = [(gerufen/a)]>
lex:)  <worden = [(werden/v)]>
lex:)  <seien = [(sein/v)]>
|
```

```
lex:)  :,/PUNC:
lex:)  <kritisierte = [(kritisieren/v), (kritisiert/a)]>
lex:)  <der = [(der/t)]>
lex:)  <Präsident = [(präsident/s)]>
lex:)  <der = [(der/t)]>
lex:)  <Gesellschaft = [(gesellschaft/s)]>
lex:)  :,/PUNC:
lex:)  <Sebastian = [(sebastian/e)]>
lex:)  <Pflugbeil|KOM = [(pflugbeil/k), (beil/s+), (pflug/s+)]>
lex:)  :(/OTHR:
lex:)  <Berlin = [(berlin/e)]>
lex:)  :)/OTHR:
lex:)  :./PUNC:
lex:)  <die = [(die/t)]>
lex:)  <Strahlenschutzverordnung|KOM = [(strahlenschutzverordnung/k), (schutz/s+), (strahl/s+),
         (verordnung/s+), (strahlen/v+)]>
lex:)  <des = [(des/t)]>
lex:)  <Bundes = [(bund/s), (bundesstaat/y)]>
lex:)  <habe = [(haben/v)]>
lex:)  <seit = [(seit/w)]>
lex:)  <Jahrzehnten = [(jahrzehnt/s)]>
lex:)  <ein = [(ein/t)]>
lex:)  <unterschätztes|KOM = [(unterschätzen/k), (schätzen/v+), (unter/w+)]>
lex:)  <Risiko = [(risiko/s)]>
lex:)  <zur = [(zur/t)]>
lex:)  <Grundlage = [(grundlage/s)]>
lex:)  :,/PUNC:
lex:)  <sagte = [(sagen/v)]>
```

```
lex:)  <Pflugbeil|KOM = [(pflugbeil/k), (beil/s+), (pflug/s+)]>
lex:)  <am|?>
lex:)  <Donnerstag = [(donnerstag/s)]>
lex:)  <in = [(in/t)]>
lex:)  <Bremen = [(bremen/e)]>
lex:)  :./PUNC:
lex:)  <Dort = [(dort/w)]>
lex:)  <beginnt = [(beginnen/v)]>
lex:)  <am|?>
lex:)  <Freitag = [(freitag/s)]>
lex:)  <der = [(der/t)]>
lex:)  <zweitägige|?>
lex:)  <international kongreß -- kongreß, international|SEQ = [(international kongreß -- kongreß,
        international/q)]>
lex:)  <internationale = [(international/a)]>
lex:)  <Kongress = [(kongreß/s)]>
lex:)  :`/OTHR:
lex:)  <Strahlenschutz|KOM = [(strahlenschutz/k), (strahlenschutzvorsorge/y), (schutz/s+), (strahl/s+),
        (strahlen/v+)]>
lex:)  <nach = [(nach/w)]>
lex:)  <der = [(der/t)]>
lex:)  <Jahrtausendwende|KOM = [(jahrtausendwende/k), (jahr 2000/y), (jahr-2000-problem/y), (jahr-
        zweitausend-problem/y), (jahrtausendende/y), (jahrtausendproblem/y), (jahrtausendwechsel/y),
        (millenium/y), (millennium/y), (millennium bug/y), (y2k/y), (year 2 kilo/y), (year 2000/y), (jahr/s+),
        (wende/s+), (tausend/w+)]>
```

```
lex:)  :"/OTHR:
lex:)  :./PUNC:
lex:)  <Er = [(er/t)]>
lex:)  <forderte|?>
lex:)  <Eine = [(ein/w)]>
lex:)  <deutlich senkung -- senkung, deutlich|SEQ = [(deutlich senkung -- senkung, deutlich/q)]>
lex:)  <deutliche = [(deutlich/a)]>
lex:)  <Senkung = [(senkung/s)]>
lex:)  <des = [(des/t)]>
lex:)  <Grenzwertes|?>
lex:)  <für = [(für/w)]>
lex:)  <beruflich = [(beruflich/a)]>
lex:)  <von|?>
lex:)  <Strahlen = [(strahl/s), (strahlen/v), (strahlverfahren/y)]>
lex:)  <betroffene = [(betroffene/s), (betroffen/a)]>
lex:)  <Personen = [(person/s)]>
lex:)  :./PUNC:
lex:)  <dpa|?>
lex:)  ://OTHR:
lex:)  <lni|?>
lex:)  <sm|?>
lex:)  <yyni|?>
lex:)  <ba|?>
lex:)  <ub|?>
lex:)  :081351/NUMS:
lex:)  <Jun|?>
lex:)  :00/NUMS:
lex:)  *EOF('txt/dpa.txt')
------------------------------------------------------------
Die Dauer der Sitzung war  0.36 sec.
```

# Indexterme, alphabetisch

anerkannt
arbeit
arbeiten
atomindustrie
beginnen
beigetragen
beitragen
bereich
berlin
beruflich
berufskrankheit
betroffen
betroffene
bremen
bund
deutlich
deutschland
donnerstag
drastisch
einer
fehleinschätzung
freitag
fällen
geben
gerufen
gesellschaft
grundlage
haben
industriefreundlich
international
jahrtausendwende
jahrzehnt
kongreß
kritisieren
kritisiert
mangelhaft
mehr
person
pflugbeil
politik
präsident
risiko
sagen
schutzbestimmung
sebastian
sein
senkung
strahl
strahlen
strahlenrisiko
strahlenschutz
strahlenschutzverordnung
unterschätzen
vorgeworfen
werden
wirtschaft
wissenschaft

## nicht erkannte Terme

-
als
am
ba
dpa
forderte
grenzwertes
im
jun
lni
sm
ub
von
yyni
zweitägige

## Synonyme

arbeitsbedingte krankheit
bundesstaat
bürgerliche wissenschaft
erwerbsarbeit
jahr 2000
jahr-2000-problem
jahr-zweitausend-problem
jahrtausendende
jahrtausendproblem
jahrtausendwechsel
kerntechnische industrie
millenium
millennium
millennium bug
politische entwicklung
politische lage
radiation hazard
staatspolitik
strahlengefährdung
strahlenschutzvorsorge
strahlungsgefährdung
strahlungsrisiko
strahlverfahren
wirtschaftsleben
wissenschaften
y2k
year 2 kilo
year 2000

## algorithmische Mehrwortbegriffe

deutlich senkung -- senkung, deutlich
drastisch fehleinschätzung -- fehleinschätzung, drastisch
international kongreß -- kongreß, international

# Projekt Milos I

## Ausgangspunkt

- unzureichender Recall im OPAC (Suche mit Titelstichwörtern < 14%)
- zu geringe Erschließungsquote (ca. 30%): Recall bei Suche mit Titelstichwörtern und Schlagwörtern < 40%
- sprachliche Probleme bei der Suche im "Basic-Index“

!

## Ziele des Einsatzes von IDX/MILOS

- Erhöhung des Recall
- Verbesserung der Suche auf Titelstichwörter
- sprachliche Vereinheitlichung (Normierung, Verdichtung) des Basic Index

## Funktionalität der Indexierung

- Standardfunktionen der wörterbuchbasierten Indexierung des Deutschen (vgl. 2.1)
- Einsprachige Indexierung für Deutsch, Englisch und Französisch
- Keine Übersetzung

# Retrievaltest MILOS I

**Basis**

- Testdatenbank mit 40.000 Titel (zufälliges Segment)
- 50 Suchanfragen, z.B.
  - Wachstum und Wirtschaft
  - Kunst der Antike
  - Haushaltswissenschaft
  - Folgen einer Scheidung für Kinder
- insg. 876 relevante Dokumente

**Ergebnisse**

- Suche mit Titelstichwörtern
  - Recall: 14%
  - Precision: 59%
  - Einheitswert: 0.84
- Suche mit Titelstichwörtern und Indexaten
  - Recall: 51%
  - Precision: 83%
  - Einheitswert: 0.46
- Suche mit Titelstichwörtern und Schlagwörtern
  - Recall: 39%
  - Precision: 83%
  - Einheitswert: 0.58

# MILOS II – Semantische Umfeldsuche

**Ausgangspunkt**

- ➢ Stichwortsuchen sind beschränkt auf das verfügbare Suchvokabular des Nutzers
- ➢ automatische Indexierung ändert dies grundsätzlich nicht, mildert allerdings sprachlich bedingte "Matching-"Probleme
- ➢ wünschenswert wäre daher eine (automatische) Möglichkeit, die Abhängigkeit zwischen Suchergebnis und Suchvokabular des Nutzers auf eine semantische Grundlage zu stellen

**Beispiel**

Nutzer sucht

*Trinkwasserbelastung*

Suche findet nicht

*Wasserverschmutzung, Trinkwasserverschmutzung, Umweltschutz, Wasserschutz etc.*

**Mögliche Lösung**

Automatische Indexierung unter Berücksichtigung **semantischer Relationen**, d.h. Synonymbeziehungen, Oberbegriffe

# MILOS II – Semantische Umfeldsuche

## Das MILOS II Verfahren

- automatische Indexierung auf grammatikalischer Ebene (vgl. Funktionalität IDX / MILOS)
- Einbeziehung von Begriffsbeziehungen der Schlagwortnormdatei
  - Synonyme
    - *Synonym → Vorzugsbenennung (= SWD-Ansetzung)*
  - Ober- und Unterbegriffe
    - *Unterbegriff → Oberbegriff*

## Der Retrievaltest

- Testpool DNB Reihe A 1991-1995 (ca. 190.000 Dokumente)
- **100 Suchfragen** (davon 50 aus Retrievaltest MILOS I) auf

Titelstichwörter Indexierungsergebnisse (MILOS)

verstichwortete RSWK-Ketten RSWK-Ketten Basic Index

# MILOS II – Ergebnisse des Retrievaltests

- verstichwortete RSWK-Ketten **doppelt** so viel relevante Dokumente wie Titelstichwörter
- Indexierungsergebnisse **dreimal** so viel relevante Dokumente wie Titelstichwörter
- **Precision**
  - Titelstichwörter 0.82
  - Indexierung 0.75
  - RSWK-Stichw. 0.95
- **Nulltreffer-Ergebnisse**
  - Titelstichwörter 15
  - Indexierung 3
  - RSWK-Stichw. 30

**Sachse, E.; Liebig, M.; Gödert, W.:**

Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt.

Köln: FH Köln, Fachbereich Bibliotheks- und Informationswesen, 1998. (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft; Bd.14)

## Ziel

Untersuchung des Sucherfolgs in einem Information-Retrieval-System auf (idealerweise) objektiver Basis

### 1. Festlegung des Dokumentenpools

- Größe
- Dokumententypus
- Homogenität
- Zufälligkeit

### 2. Festlegung von Suchanfragen

- Anzahl der Fragen
- Fragetypus
- thematische Streuung

### 3. Festlegung von Suchverfahren

- Durchführung: Laie vs. Experte
- Umsetzung der Suchanfragen in eine Retrievalstrategie
  - formal: Syntax von Thema und Frage
  - inhaltlich: Umsetzung des Inhalts der Suchanfrage

## 4. Festlegung von Kriterien für Trefferdokumente / Relevanzkriterien

- Welche gefundenen Dokumente sind relevant, welche nicht?
- Wer entscheidet das?

## 5. Berechnung von (objektiven) Maßzahlen

## 6. Interpretation der Ergebnisse

- Was ist gut?
- Warum?

$$\textbf{Recall} = \frac{\text{gefundene relevante Dokumente}}{\text{alle relevanten Dokumente}}$$

$$\textbf{Precision} = \frac{\text{gefundene relevante Dokumente}}{\text{alle gefundenen Dokumente}}$$

# Automatische Indexierung

# Statistische Verfahren

**Allgemeine Überlegungen:**

- Es besteht ein Zusammenhang zwischen der Auftretenshäufigkeit von Wörtern und deren Bedeutung für das Retrieval.
- Wichtig sind solche Wörter, die
  - Dokumente hinreichend signifikant vertreten und gleichzeitig
  - von nicht-relevanten Dokumenten trennen.

**Verteilung der Worthäufigkeit in Textkorpora: "Zipf's Law"**

Worthäufigkeit * Häufigkeitsrang = Konstante

Worthäufigkeit: Auftretenshäufigkeit eines Wortes/Kollektion

Häufigkeitsrang: Position im Ranking nach Häufigkeit

| **Beispiel:** | |
|---|---|
| 1. Häufigstes Wort | 10.000 |
| 2. Zweithäuf. Wort | 5.000 |
| 3. Dritthäuf. Wort | 3.300 |
| 10.000. Zehn... | 1 |

**Wortverteilung in den Kollektionen von TREC-1 (1993)**

| Quelle | WSJ | AP | ZIFF | FR | DOE |
|---|---|---|---|---|---|
| Größe in MB | 295 | 266 | 251 | 258 | 190 |
| Mittelwert: Wörter/DS | 182 | 353 | 181 | 313 | 82 |
| Verschiedene Wörter | 156.000 | 198.000 | 174.000 | 126.000 | 186.000 |
| Einmaliges Auftreten | 65.000 | 90.000 | 86.000 | 59.000 | 96.000 |
| Auftreten > 1 | 199 | 174 | 165 | 106 | 159 |

**Vermutungen**

- hochfrequente Wörter sind schlechte Suchbegriffe
- niedrigfrequente Wörter sind schlechte Suchbegriffe, weil sie wahrscheinlich nicht zum Vokabular des Nutzers gehören und/oder autorenspezifisch sind

**(1) Einfache Termhäufigkeit (TF)**

*"Je häufiger ein Term in einem Dokument vorkommt, umso wichtiger ist er für dieses Dokument."*

**Termhäufigkeit = Häufigkeit Term je Dokument**

**(2) Relative Termhäufigkeit (WDF)**

*"Die einfache Termhäufigkeit bevorzugt lange Dokumente, interessant ist also die Termhäufigkeit in Relation zur Länge des Dokuments."*

***WDF = Häufigkeit Term je Dokument / Gesamtzahl Terme im Dokument***

**(3) Dokumenthäufigkeit (DF)**

*"Je weniger Dokumente es zu einem Term gibt, umso wichtiger ist der Term."*

**Dokumenthäufigkeit = Häufigkeit Dokumente je Term**

**(4) Inverse Dokumenthäufigkeit (IDF)**

*"Terme, die in einigen Dokumente häufig, insgesamt aber nicht so häufig vorkommen, sind wichtig."*

**TF*IDF = Termhäufigkeit bzw. WDF / Dokumenthäufigkeit**

# Vergleich von Termgewichtungsverfahren

**5 Dokumente aus einer Kollektion von 10.000**

d1 = Anwendung des Prinzips Thesaurus für das Retrieval im OPAC

d2 = Zusammenhang zwischen Thesaurus und Klassifikation

d3 = Klassifikation und OPAC: verbesserter Sucherfolg durch Einsatz einer Klassifikation im Retrieval

d4 = Thesaurus für die Physik und Thesaurus für die physikalische Chemie

d5 = Klassifikation für die Chemie

**Anzahl der Dokumente mit den Suchtermen**

| | |
|---|---|
| Anwendung | 2000 |
| Chemie | 200 |
| Einsatz | 100 |
| Klassifikation | 100 |
| OPAC | 600 |
| Physik | 300 |
| Prinzip | 1500 |
| Retrieval | 400 |
| Sucherfolg | 50 |
| Thesaurus | 200 |
| Zusammenhang | 3000 |

- Berechnen Sie die relative Termhäufigkeit WDF und die inverse Dokumenthäufigkeit IDF für alle Suchterme (nur Substantive) in den Dokumenten.
- Beispiel:
  d1 = Anwendung (Gewicht), Prinzip (Gewicht), ...
- Berechnen Sie die Retrievalergebnisse für folgende Suchanfragen:
  - **Thesaurus im Retrieval**
  - **Klassifikation in der Chemie**
- Diskutieren Sie den Nutzen der Gewichtung im Hinblick auf das Retrieval, insb. das Ranking

# Ergebnisse

**Ergebnisse IDF für Suchfrage "Thesaurus and Retrieval"**

| | | | |
|---|---|---|---|
| d1 | 1/200 + 1/400 | = | 3/400 (2) |
| d2 | 1/200 | = | 1/200 (3) |
| d3 | 1/400 | = | 1/400 (4) |
| d4 | 1/100 | = | 1/100 (1) |
| d5 | -- | = | -- |

**Ergebnisse IDF für Suchfrage "Klassifikation and Chemie"**

| | | | |
|---|---|---|---|
| d1 | -- | = | -- |
| d2 | 1/100 | = | 1/100 (3) |
| d3 | 1/50 | = | 1/50 (1) |
| d4 | 1/200 | = | 1/200 (4) |
| d5 | 1/100 + 1/200 | = | 3/200 (2) |

# Ergebnisse

**Ergebnisse WDF für Suchfrage "Thesaurus and Retrieval"**

| | | | |
|---|---|---|---|
| d1 | 1/5 + 1/5 | = | 2/5 (2) |
| d2 | 1/3 | = | 1/3 (3) |
| d3 | 1/6 | = | 1/6 (4) |
| d4 | 1/2 | = | 1/2 (1) |
| d5 | -- | = | -- |

**Ergebnisse WDF für Suchfrage "Klassifikation and Chemie"**

| | | | |
|---|---|---|---|
| d1 | -- | = | -- |
| d2 | 1/3 | = | 1/3 (2) |
| d3 | 1/3 | = | 1/3 (2) |
| d4 | 1/4 | = | 1/4 (3) |
| d5 | 1/2 + 1/2 | = | 1 (1) |

# AIR/PHYS

**Umgebung**

Fachdatenbank PHYS (inzw. Bestandteil von INSPEC) mit **englisch-sprachiger** Erschließung durch **normiertes Vokabular** (Deskriptoren) und **Abstracts**

**Ziel von AIR/PHYS**

automatische Indexierung der Dokumente mit **Deskriptoren** des PHYS-Thesaurus

**Realisierung**

1. **statistische Auswertung** der intellektuell erschlossenen Dokumente: v.a. Untersuchung der Beziehung

   *Term* ⇨**z** ⇨ *Deskriptor*,

   wobei **z** ein Maß für die Wahrscheinlichkeit ist, mit der ein *Deskriptor* einem Dokument (intellektuell) zugeteilt ist, wenn *Term* im Dokument vorhanden ist:

$$z = \frac{h(t,s)}{f(t)}$$

   **h(t,s)** = Anzahl der Dokumente, in denen **Term t** vorkommt und **Deskriptor s** vergeben wurde

   **f(t)** = Anzahl der Dokumente, in denen **Term t** vorkommt

1. (automatischer) **Aufbau eines Indexierungswörterbuchs** unter Ausnutzung der Gewichte aus 1., echter Thesaurusrelationen (Synonym) und Deskriptor-Deskriptor-Relationen als gewichtetes Maß für das gemeinsame Auftreten von Deskriptoren
2. **Automatische Indexierung** in zwei Phasen
    - **Rohindexierung** mit regel- und lexikonbasierter Textanalyse und statistischer Relationierung
    - **Abgestimmte Indexierung** unter Einbeziehung von Deskriptor-Deskriptor-Relationen

**Pilotanwendung AIR/PHYS**

- Wörterbuchaufbau auf der Basis von 400.000 intellektuell erschlossenen Dokumente
    - 20.000 Deskriptoren, 190.000 Wörter
    - 350.000 statistische Regeln mit z > 0,3
    - 70.000 Synonym-Relationen
    - 200.000 Deskriptor-Deskriptor-Relationen
- Erschließung von 10.000 Dokumenten / Monat
- Zuteilung von im Schnitt 12 Deskriptoren je Dokument
- intellektuelle Nachbearbeitung mit durchschnittlich einem Drittel Korrekturbedarf, d.h. **semi-automatisches Verfahren**

# Ergebnisse der AIR/PHYS-Indexierung

**Retrievaltest** mit 15.000 Dokumenten und 300 (Original-)Fragen

- automatische Indexierung

  Precision: 0.46
  Recall: 0.57

- intellektuelle Indexierung

  Precision: 0.53
  Recall: 0.51

**intellektuelle Bewertung** der Erschließungsqualität durch Experten

- 1/3 intellektuelle Erschließung besser
- 1/3 automatische Indexierung besser
- 1/3 qualitativ gleichwertig

**Literatur**

**Knorz, Gerhard**: Automatische Indexierung. In: Wissensrepräsentation und Information Retrieval. Potsdam 1994. S. 138-198.

**Nohr, Holger**: Automatische Indexierung: Einführung in betriebliche Verfahren, Systeme und Anwendungen. Potsdam 2001. S.71-77.

# Das Projekt KASCADE

**Ziele**

- Anreicherung von Titeldaten aus dem Fach Jura
- Entwicklung einer selektiven automatischen Indexierung zur gewichteten Extraktion von Deskriptoren (SELIX)
- Entwicklung einer zuverlässigen Erkennung für Themen-Aspekt-Beziehungen in Mehrwortgruppen (THEAS)
- Durchführung eines umfangreichen Retrievaltests

*Anreicherung der Titeldaten*

- Scanning von Inhaltsverzeichnissen von ca. 3.000 Titeln aus dem Bestand Jura zur Verbreiterung der Indexierungsbasis (Abstracts: zu selten; Sachregister: problematisch)
- OCR mit *newsWorks* und MILOS-Rechtschreibkontrolle

*Automatische Indexierung mit SELIX*

- MILOS-Indexierung mit Grundformermittlung und Dekomposition für Sachtitel, Schlagwörter und Volltexte der Inhaltsverzeichnisse
- SELIX-Gewichtungsindexierung

# Die KASCADE-Gewichtung

- ➢ MILOS II-Indexierung der SELIX-Deskriptoren zur Ermittlung semantischer (SWD-) Relationen (nur Synonym-Relationen)
- ➢ zusätzliche MILOS II-Indexierung von Sachtitel, Schlagwörtern und Volltexten der Inhaltsverzeichnisse mit:
  - Grundformermittlung, Dekomposition, Derivation, Synonymen

**Gewichtungsfunktion Salton**

HfkImD(g,d) * log (nDok(g) / nAnzDok)

[Termhäufigkeit * log (Dokumenthäufigkeit / Dokumentenzahl), vgl. IDF]

**Robertson**

((K 1 + 1) * HfkImD(g,d) / (K + HfkImD(g,d)))
* log((nAnzDok – nDok(g) + 0.5) / (nDok(g) + 0.5))

# Die KASCADE-Gewichtung

## Kollektionsgewicht nG1

nG1 ermittelt, ob eine Grundform für eine **Dokumentenkollektion** als Indexterm geeignet ist (für alle Dokumente (einer Kollektion) gleich).

$$nG1(g) = 1 - nDok(g) / E(nDok(g))$$

(für: $nDok(g) < E(nDok(g))$; 0 sonst)

mit

$$E(nDok(g)) = nAnzDok * (1 - \exp(-\lambda))$$

wenn eine Poisson Zufallsverteilung angenommen wird:

$$P(i) = \exp(-\lambda) * (\lambda^i / i!) \qquad \lambda = nColl(g) / nAnzDok$$

## Dokumentgewicht nG2

nG2 ermittelt, ob eine Grundform für ein **Dokument** als Indexterm wichtig ist.

$$nG2(g,d) = (p(1) * 1 + ... + p(HfkImD(g,d)) * HfkImD(g,d)) / \lambda$$

mit

$$P(i) = \exp(-\lambda) * (\lambda^i / i!)$$

$$\lambda = nColl * (nDokLen/nColLen)$$

# Die KASCADE-Gewichtung

**Längengewicht nG3**

nG3 bevorteilt längere Wörter im Gewichtungsverfahren.(unabhängig von Dokument und Kollektion)

$$nG3(g) = \log(nGruLen(g)) / 4$$

**Gewichtungsfunktion**

$$nG = F1 * nG1 + F2 * nG2 + F3 * nG3$$

wobei F1-F3 frei wählbar sind (Standard: 1)

**Hüther, H.**: *Selix im DFG-Projekt Kascade*. In: Knowledge Management und Kommunikationssysteme: Proceedings des 6. Internationalen Symposiums für Informationswissenschaft (ISI '98) Prag, 3.-7. November 1998 /

Hochschulverband für Informationswissenschaft (HI) e.V. Konstanz ; Fachrichtung Informationswissenschaft der Universität des Saarlandes, Saarbrücken. Hrsg.: Harald H. Zimmermann u. Volker Schramm. Konstanz: UVK Universitätsverlag, 1998. S.397-403.

(Schriften zur Informationswissenschaft; Bd.34)

# Der Kascade-Retrievaltest

**Rahmenbedingungen**

- 3.000 Referenzdatensätze aus dem Fach Jura
- alle angereichert um Inhaltsverzeichnisse im Volltext
- 60 von Juristen formulierte Suchthemen
- Testdurchführung durch Projektmitarbeiter
- Relevanzbewertung durch Juristen
- Recall-Berechnung nach Pooling-Methode

**Besonderheiten bei den Suchthemen**

- breite thematische Streuung – speziell neben allgemein
- viele Komposita und Mehrwortbegriffe
- viele komplexe Themen, d.h. Themenverknüpfungen
- nur 15% Einwort-Suchthemen (mit nur einem Nichtkompositum)

# Ergebnisse

| | Mittelwerte von Recall und | Precision | Null-Treffer-Suchen |
|---|---|---|---|
| Titel und Deskriptor (automatisch indexiert) | **0.06** | **0.98** | **42** |
| Titel, Deskriptor, Inhaltsverz. (nicht automatisch indexiert) | **0.54** | **0.75** | **7** |
| Titel, Deskriptor, Inhaltsverz. (automatisch indexiert) | **0.92** | **0.70** | **4** |

**Lohmann, Hartmut**: *KASCADE: Dokumentanreicherung und automatische Inhaltserschließung: Projektbericht und Ergebnisse des Retrievaltests*. Düsseldorf: Universitäts- und Landesbibliothek, 2000. 109 S.
(Schriften der Universitäts- und Landesbibliothek Düsseldorf; 31)

# Automatische Klassifizierung

Ziel: **Strukturierung großer Dokumentmengen**

Zwei Ansätze:

- **Automatisches Klassifizieren**
  als Zuweisen von Dokumenten in vorgegebene Themen
- **Clustering**
  als Unterteilung einer Dokumentkollektion in Gruppen ähnlicher Dokumente (Cluster)

## Automatisches Klassifizieren

**Ausgangspunkt**

Systematisch geordnete Themen / Klassifikation

**Ziel**

Zuordnung aller Dokumente einer Kollektion zu den Themen der Ordnung / Klassen der Klassifikation

# Automatische Klassifizierung

## Verfahren

Erstellen einer Testkollektion, d.h. intellektuelle Zuweisung von Dokumenten zu den Themen / Klassen

Analyse der Termbeziehungen in den Dokumenten einer Klasse, z.B. auf der Basis einer **Dokument-Term-Matrix** der gewichteten Terme:

| | **Term1** | **Term2** | **Term3** | **Term4** | **Term5** | **Term6** | **Term7** | **Term8** |
|---|---|---|---|---|---|---|---|---|
| **Dok 1** | 0 | 4 | 0 | 0 | 0 | 2 | 1 | 3 |
| **Dok 2** | 3 | 1 | 4 | 3 | 1 | 2 | 0 | 1 |
| **Dok 3** | 3 | 0 | 0 | 0 | 3 | 0 | 3 | 0 |
| **Dok 4** | 0 | 1 | 0 | 3 | 0 | 0 | 2 | 0 |
| **Dok 5** | 2 | 2 | 2 | 3 | 1 | 4 | 0 | 2 |

# Automatische Klassifizierung

- ➢ Ermittlung der häufigsten gemeinsamen Terme einer Klasse
- ➢ Ermittlung der Häufigkeit dieser Terme in anderen Klassen
- ➢ Zuweisung der Terme zur Klasse, falls Terme in der Klasse häufig, in anderen Klassen jedoch selten sind

**Ergebnis**

Zuordnung von Termen zu Klassen

**Klassifikationsverfahren**

- ➢ Festlegung der Bedingungen, die zur Zuweisung eines Dokuments zu einer Klasse führen:
  - wie viele Terme einer Klasse müssen mindestens im Dokument enthalten sein
  - welche Gewichte müssen diese haben
- ➢ Termgewichtung für neue Dokumente
- ➢ Anwendung der Regeln
- ➢ Zuordnung eines Dokuments zu einer Klasse

# Clustering

**Ausgangspunkt**

unstrukturierte, in der Regel sehr große Dokumentkollektion

**Ziel**

Strukturierung der Kollektion durch Ermittlung von Gruppen ähnlicher Dokumente

**Verfahren**

Berechnung der Ähnlichkeit von Dokumenten

- durch Analyse der Beziehungen zwischen Dokumenten und den in ihnen enthaltenen Termen
- und Festlegung eines Clustering-Algorithmus‘ für die Zuweisung von Dokumenten zu Clustern

# Clustering

**Dokument-Term-Matrix**,

d.h. welche Dokumente enthalten welche Terme mit welchem Gewicht

| | **Term1** | **Term2** | **Term3** | **Term4** | **Term5** | **Term6** | **Term7** | **Term8** |
|---|---|---|---|---|---|---|---|---|
| **Dok1** | 0 | 4 | 0 | 0 | 0 | 2 | 1 | 3 |
| **Dok2** | 3 | 1 | 4 | 3 | 1 | 2 | 0 | 1 |
| **Dok3** | 3 | 0 | 0 | 0 | 3 | 0 | 3 | 0 |
| **Dok4** | 0 | 1 | 0 | 3 | 0 | 0 | 2 | 0 |
| **Dok5** | 2 | 2 | 2 | 3 | 1 | 4 | 0 | 2 |

# Clustering

Erzeugung einer **Dokument-Dokument-Matrix** durch Berechnung der Skalarprodukte von jeweils zwei Dokumentvektoren

| | **Dok1** | **Dok2** | **Dok3** | **Dok4** | **Dok5** |
|---|---|---|---|---|---|
| **Dok1** | | 11 | 3 | 6 | 22 |
| **Dok2** | 11 | | 12 | 10 | 36 |
| **Dok3** | 3 | 12 | | 6 | 9 |
| **Dok4** | 6 | 10 | 6 | | 11 |
| **Dok5** | 22 | 36 | 9 | 11 | |

# Clustering

Erzeugung einer **Dokument-Beziehungs-Matrix** durch Festlegung eines Schwellenwertes (hier: 10)

| | **Dok1** | **Dok2** | **Dok3** | **Dok4** | **Dok5** |
|---|---|---|---|---|---|
| **Dok1** | | 1 | 0 | 0 | 1 |
| **Dok2** | 1 | | 1 | 1 | 1 |
| **Dok3** | 0 | 1 | | 0 | 0 |
| **Dok4** | 0 | 1 | 0 | | 1 |
| **Dok5** | 1 | 1 | 0 | 1 | |

# Clustering

Anwendung eines **Clusteralgorithmus‘** zur Verteilung der Dokumente auf Cluster

**Clusteralgorithmen**

- **Cliquen-Algorithmus**
  alle Dokumente eines Clusters sind allen anderen Dokumenten des Clusters ähnlich; Dokumente in einem Cluster haben die engstmögliche Beziehung zueinander – Dokumente eines Clusters repräsentieren ein Thema (Topic)
- **Single-Link-Algorithmus**
  jedes Dokument eines Clusters ist mindestens einem Dokument des Clusters ähnlich; Dokumente eines Clusters haben schwache Beziehung zueinander – Dokumente eines Clusters repräsentieren keine Themen
- **Varianten** zwischen beiden Extremen

# Clustering

**Spielarten**

(1) Verwendung von Startclustern und Berechnung von **Zentroiden**

- Festlegung von Clustern und beliebige Zuweisung von Dokumenten zu Clustern
- Berechnung eines Zentroids (d.h. eines Mittelwerts aller Dokumente eines Clusters)
- Berechnung der Ähnlichkeit zwischen den Dokumenten in den Clustern und den Zentroiden der Cluster und Neuverteilung der Dokumente in die Cluster
- Durchführung des Verfahrens bis zu stabilen Clustern

(2) **Hierarchisches Clustering**, z.B. durch

- iteratives Clustern von erzeugten Clustern bis hin zum einzelnen Dokument (Top-down)
- Berechnung von Zentroiden für die Cluster und Clustering der Zentroide (erzeugt erste hierarchisch höhere Ebene; Bottom-up)
- Fortführung des Prozesses bis zur gewünschten Hierarchie

# Clustering

## Nutzen von Clustering im Information Retrieval

### Termclustering

Clustering von **Termen** einer Kollektion erzeugt Mengen ähnlicher Begriffe, die für die automatische Erstellung thesaurus-ähnlicher Werkzeuge verwendet werden können:

- Ausweitung der Suche durch Einbeziehung ähnlicher Begriffe;
- Verlassen der strengen Matching-Bedingungen im Zeichenketten-Retrieval;
- Angleichung von Such- und Autorensprache;
- Visualisierung von Begriffsbeziehungen.

### Dokumentclustering

Clustering von Dokumenten einer Kollektion erzeugt Mengen ähnlicher Dokumente, die für die Suche verwendet werden können:

- Ausweitung der Suche auf ähnliche Dokumente;
- Strukturierung von Treffermengen (NorthernLight-Prinzip);
- Visualisierung von Dokumentbeziehungen in Suchergebnissen;
- Verlassen der strengen Matching-Bedingungen im Zeichenketten-Retrieval;
- Relevance Feedback

# Literatur

**Kowalski, Gerald J.; Maybury, Mark T.:** *Information Storage and Retrieval Systems: Theory and Implementation*. Second Edition. Boston 2000.

Hier: Kapitel 6: *Document and Term Clustering*, S. 139-163.

# Literatur

www.indexierung-retrieval.de


**Gödert, Winfried; Lepsky, Klaus; Nagelschmidt, Matthias:** Informationserschließung und Automatisches Indexieren: Ein Lehr- und Arbeitsbuch. Berlin 2012.

**Lepsky, Klaus:** Automatische Indexierung. In: Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -Praxis. Hrsg. von R. Kuhlen, W. Semar, und D. Strauch, S. 272–85. Berlin: De Gruyter, 2013.

**Knorz, Gerhard:** Automatische Indexierung. In: Wissensrepräsentation und Information-Retrieval. Universität Potsdam 1994. S. 138-198.

**Lepsky, Klaus:** Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen. Köln 1994.
(Kölner Arbeiten zum Bibliotheks- und Dokumentationswesen ; Heft 18)

**Lepsky, Klaus:** Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. Düsseldorf: Universitäts- und Landesbibliothek, 1996. In: Zukunft der Sacherschließung im OPAC: Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995. Hrsg.: E. Niggemann u. K. Lepsky. S.13-36.
(Schriften der Universitäts- und Landesbibliothek Düsseldorf; Bd.25)

**Lepsky, Klaus:** Automatische Indexierung zur Erschließung deutschsprachiger Dokumente. In: nfd Information - Wissenschaft und Praxis. 50(1999), H.6, S.325-330.

**Lepsky, Klaus; Vorhauer, John:** Lingo : ein open source System für die Automatische Indexierung deutschsprachiger Dokumente. In: ABI Technik (2006), Nr. 1, S. 18-28.

**Nohr, Holger:** Grundlagen der automatische Indexierung. Ein Lehrbuch. Berlin 2005.