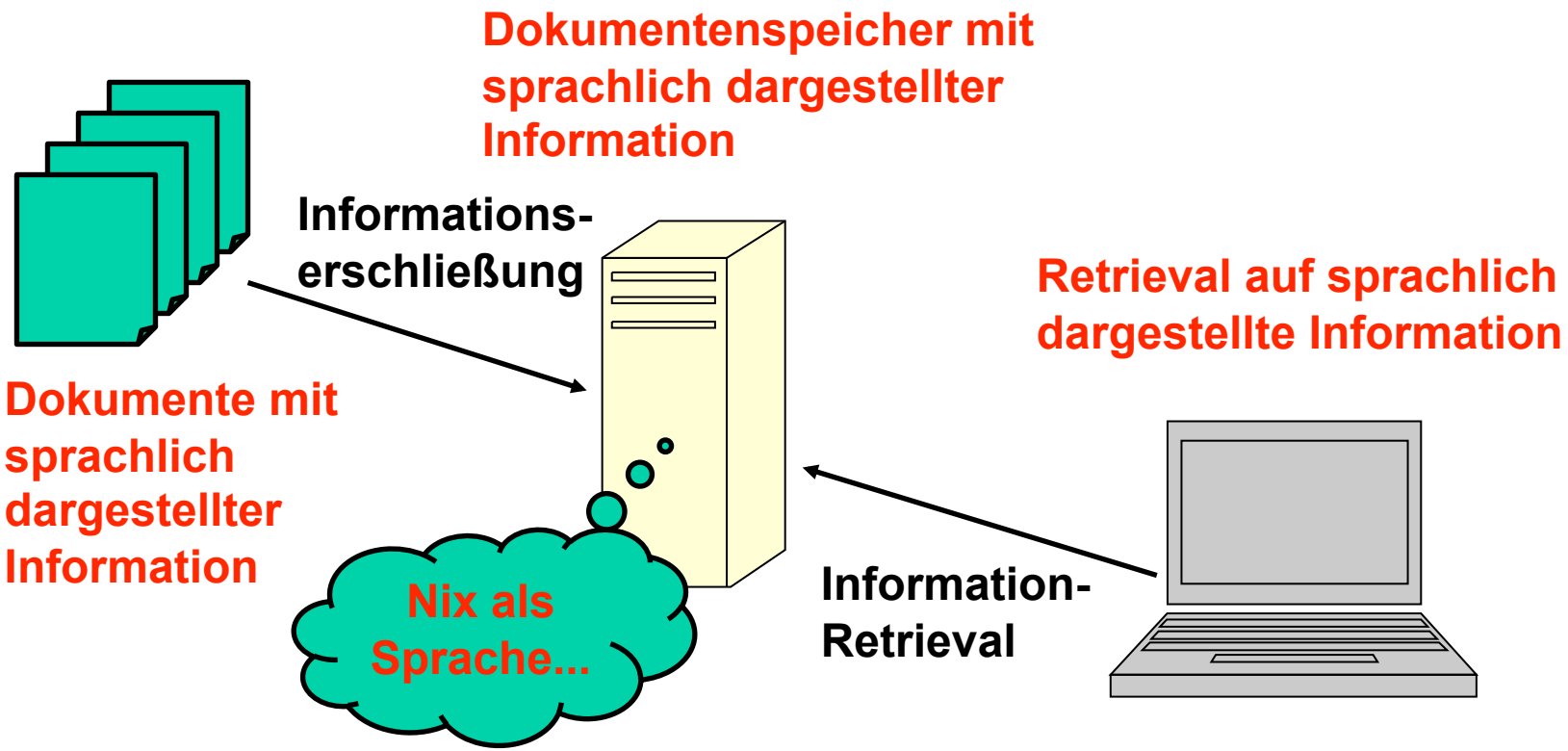
Sprachengineering

Grundlagen und Methoden sprachverarbeitender Verfahren

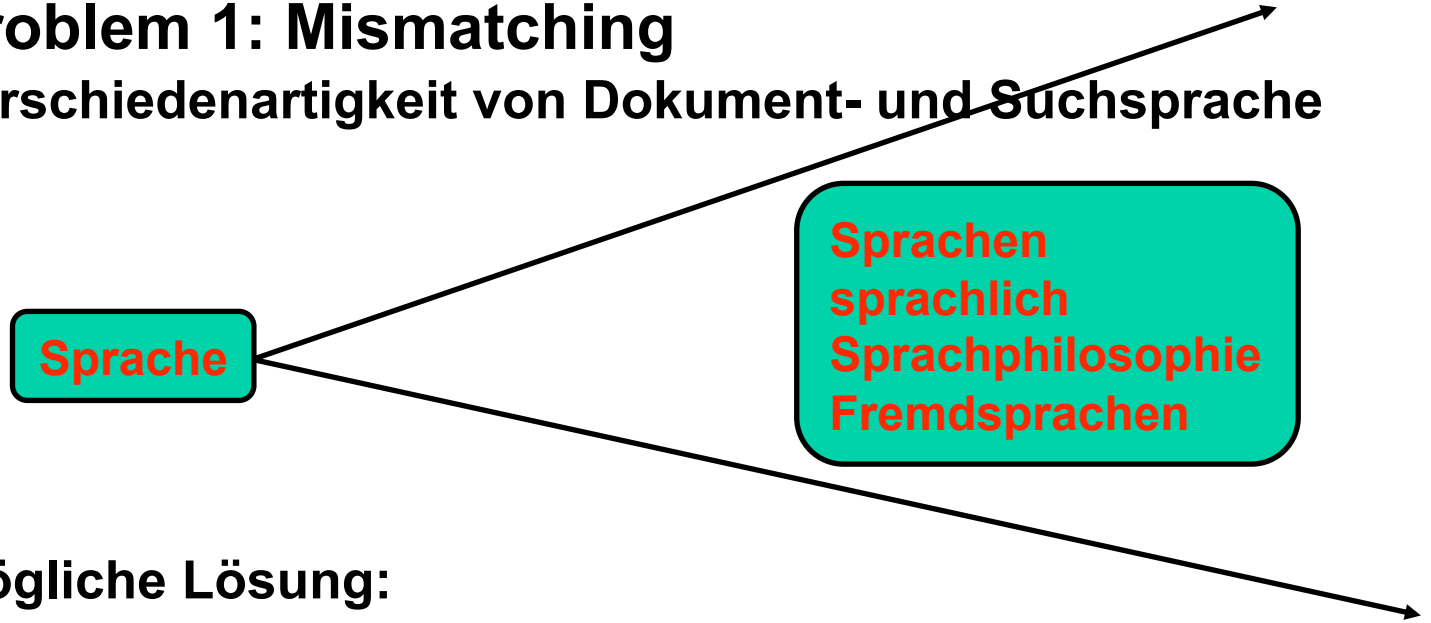
1. Einführung: Sprache und Information

Informationssuche ist in der Regel Suche nach oder in Text bzw. Sprache.

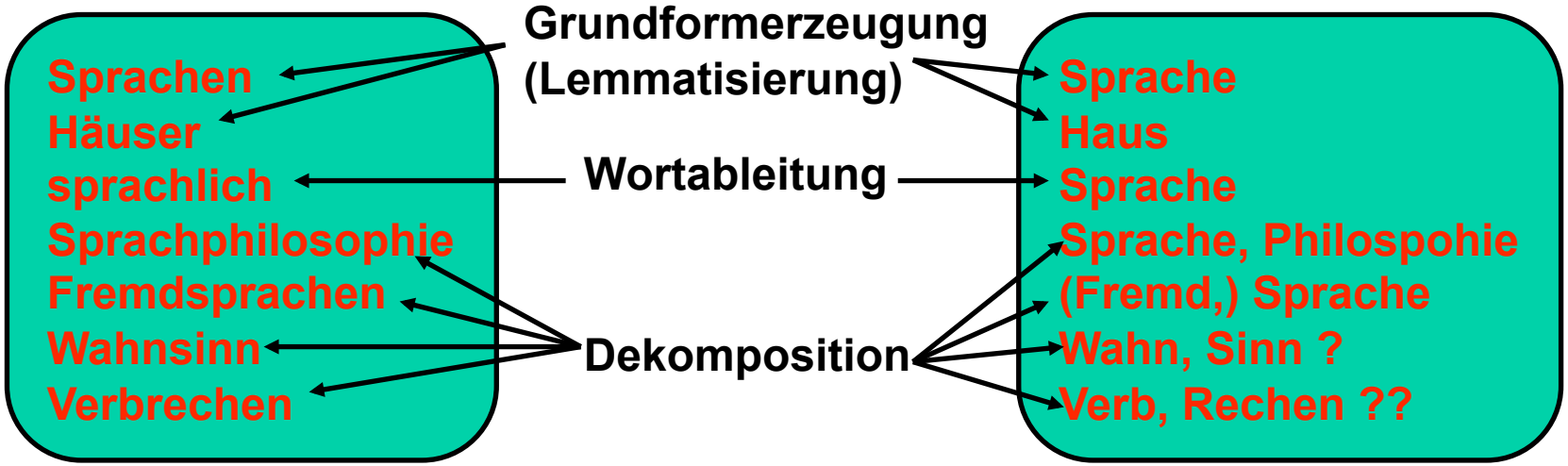


Problem 1: Mismatching

Verschiedenartigkeit von Dokument- und Suchsprache

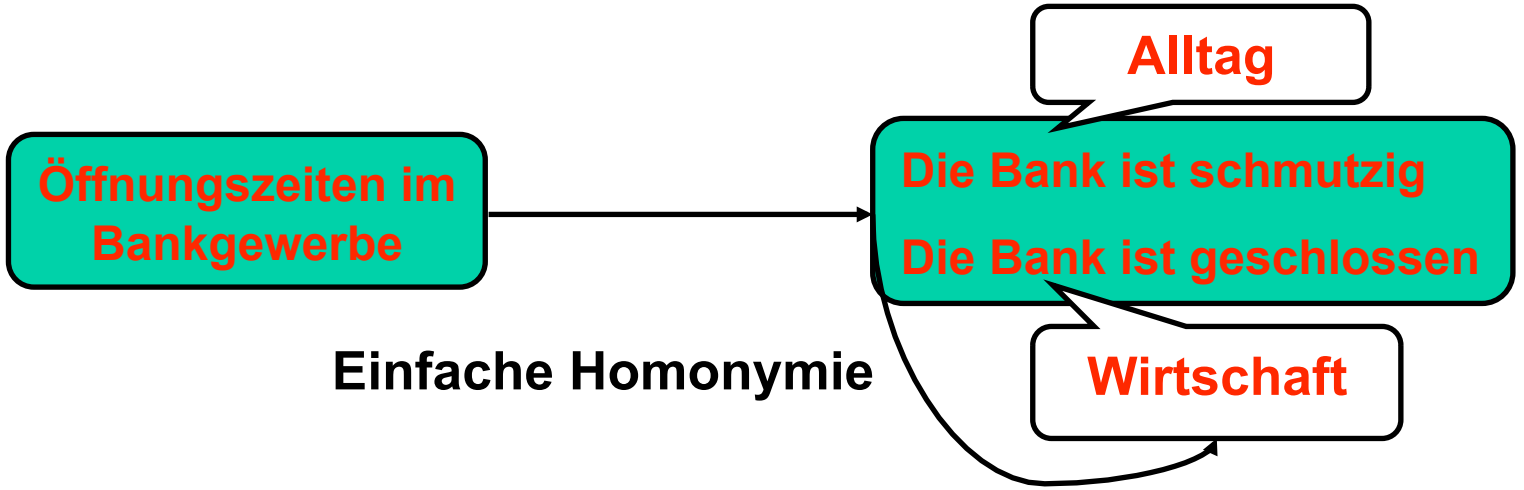


Mögliche Lösung:
Einsatz einer morphologischen Komponente



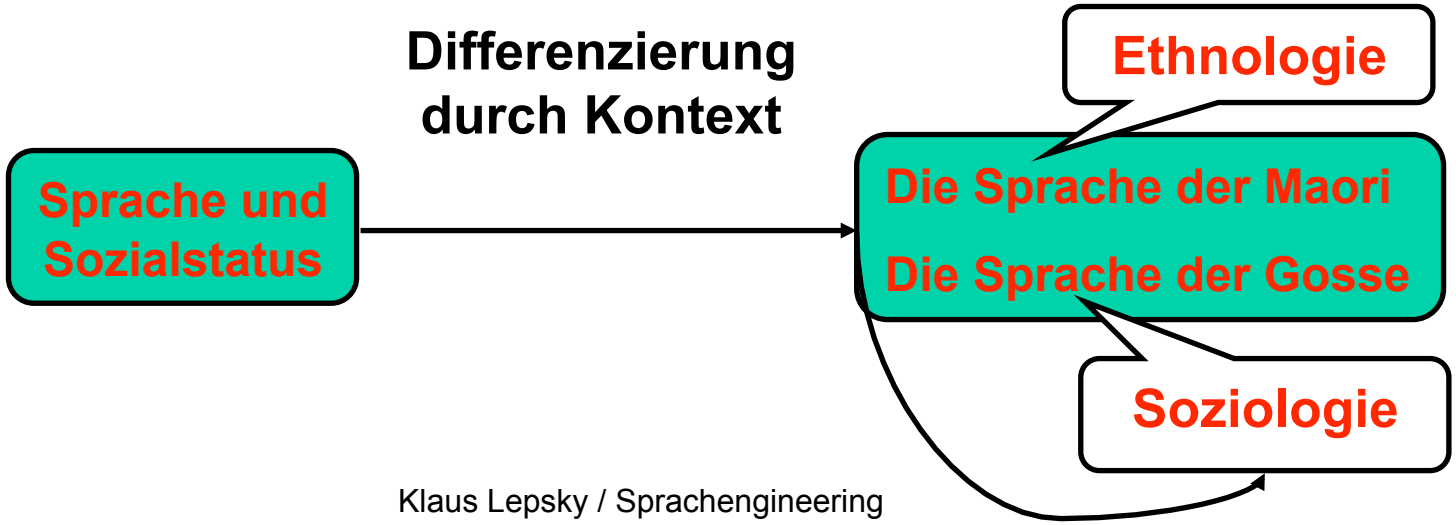
Problem 2: Bedeutungsdifferenzierung

Trennung von sprachlicher Form und Bedeutung



Einfache Homonymie

Differenzierung durch Kontext



Mögliche Lösung:
Einsatz einer kontextsensitiven Sprachanalyse

1. Partielles oder vollständiges Parsing

Identifizierung von Satzzusammenhängen

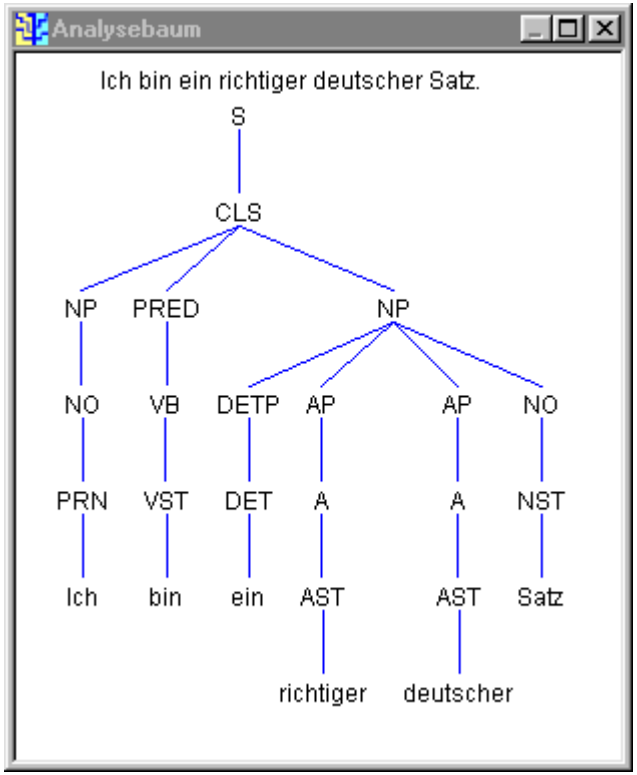
Die Bank betrieb seit langem schmutzige Geschäfte

Identifizierung von "Mehrwortgruppen" bzw. "Themen-Aspekt-Beziehungen"

progressives Steuermodell

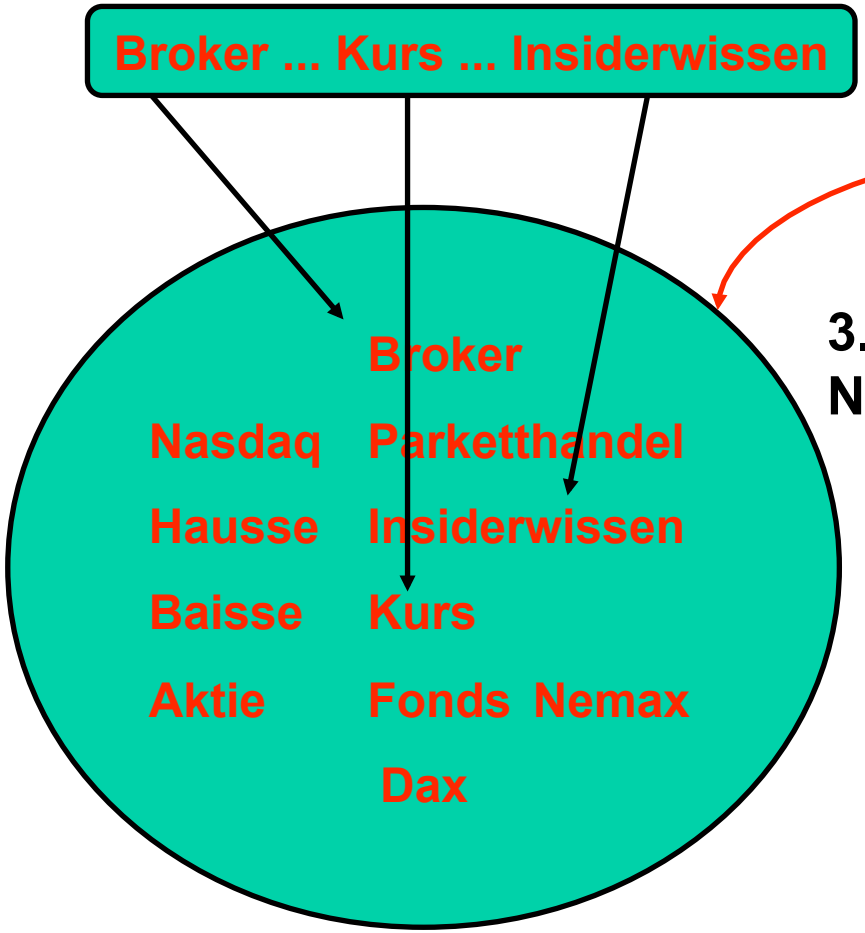
Lehrbuch der Psychologie

Information und Kommunikation



2. Einsatz einer statistischen Komponente zur Gewichtung

Identifizierung von thematischen Beziehungen



3. Einsatz eines semantischen Netzes

ermöglicht zuteilende automatische Inhaltserschließung

löst durch Synonymie und Homonymie hervorgerufene Probleme

Bedingungen und Grenzen für den Einsatz angewandter Computerlinguistik in der Informationserschließung

1. **Die Textbasis der Dokumente muss ausreichend sein**

   Volltexte → Inhaltsverzeichnisse → Abstracts → Titel

2. **Der für die Verarbeitung verfügbare Text muss ausreichende Aussagen über den Inhalt der Dokumente machen**

3. **Die Datenbasis für statistische und/oder zuteilende Verfahren muss hinreichend homogen sein**

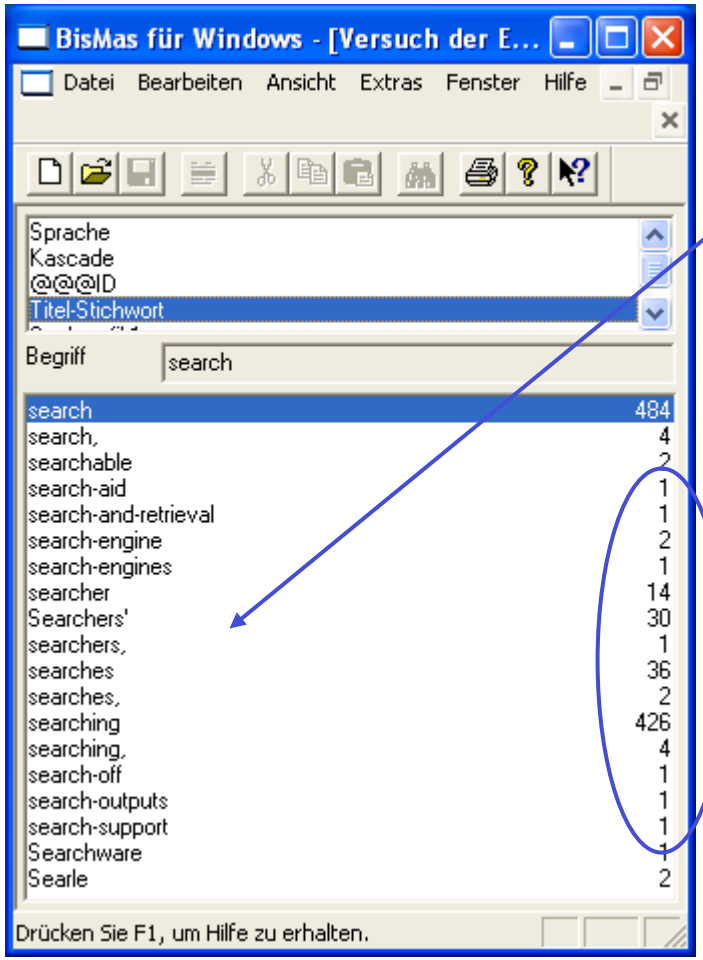
Allgemein sind die Grenzen automatischer Sprachverarbeitung dort erreicht, wo die Intelligenz beginnt.

Themen

1. Einführung: Sprache und Information ✓
2. Sprachverarbeitung im Information Retrieval
    - 2.1 Stemming des Englischen
    - 2.2 Grundformreduktion des Deutschen
    - 2.3 Kompositumzerlegung
    - 2.4 Extraktion von Phrasen
    - 2.5 Semantisches Umfeld
    - 2.6 Automatische Indexierung
3. Abstracting und Summarizing
4. Textanalyse
    - 4.1 Formale Grammatiken
    - 4.2 Parsing
5. Automatische Übersetzung
6. Literatur

2. Sprachverarbeitung im Information Retrieval

2.1 Stemming des Englischen

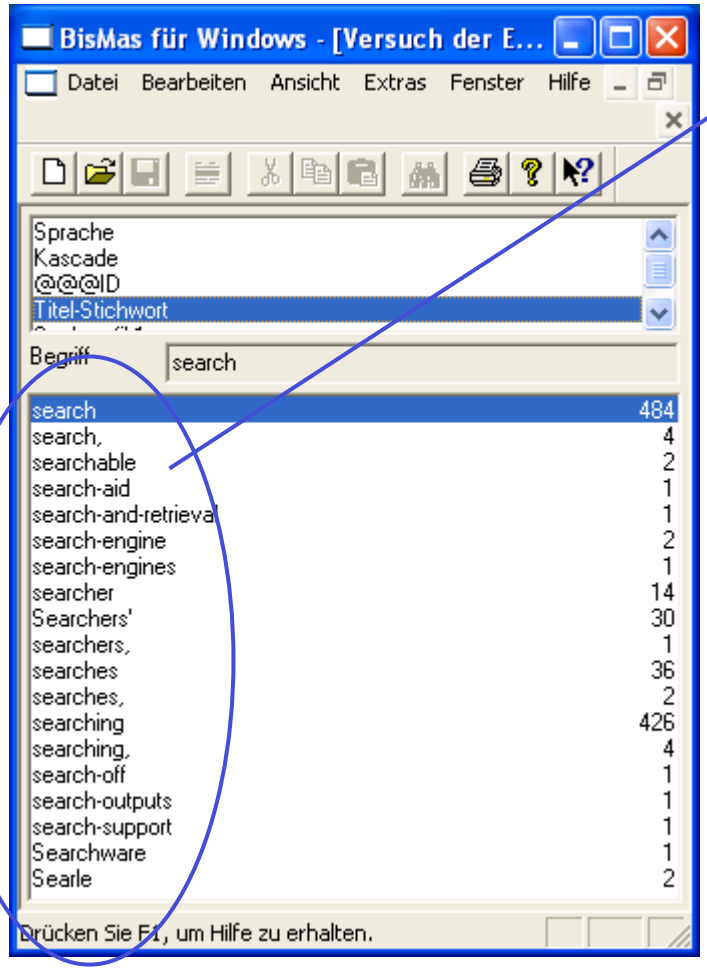


Stichwortindex einer Datenbank

Verteilung der Treffer auf unterschiedliche Wortformen

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

  Maus - **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

- **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** macht nur dann Sinn, 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 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.

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; Flexionsendungen 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*]

Ein einfaches Stemming-Regelwerk

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%!

Übung 1

? Testen Sie das Regelwerk für folgende Beispiele; welche Regeln werden jeweils angewandt:

algorithms, associated, indexings, inverted, ladies', mother's, properties, satisfied, searches, using

? Entwerfen Sie einen Stemming-Algorithmus für Pluralendungen deutscher Substantive.

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 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

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

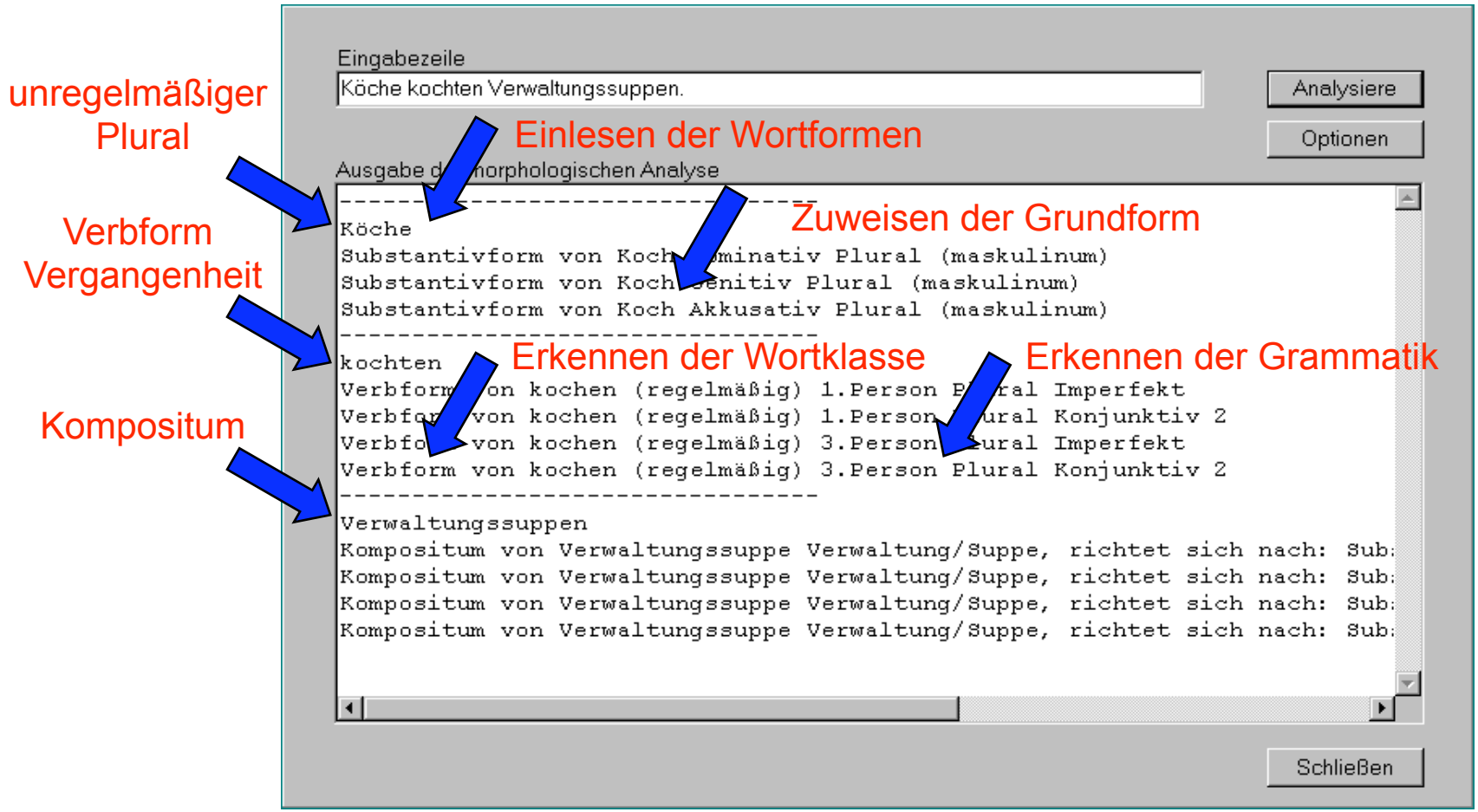
und, oder (*koordinierend*)

weil, nachdem (*subordinierend*)

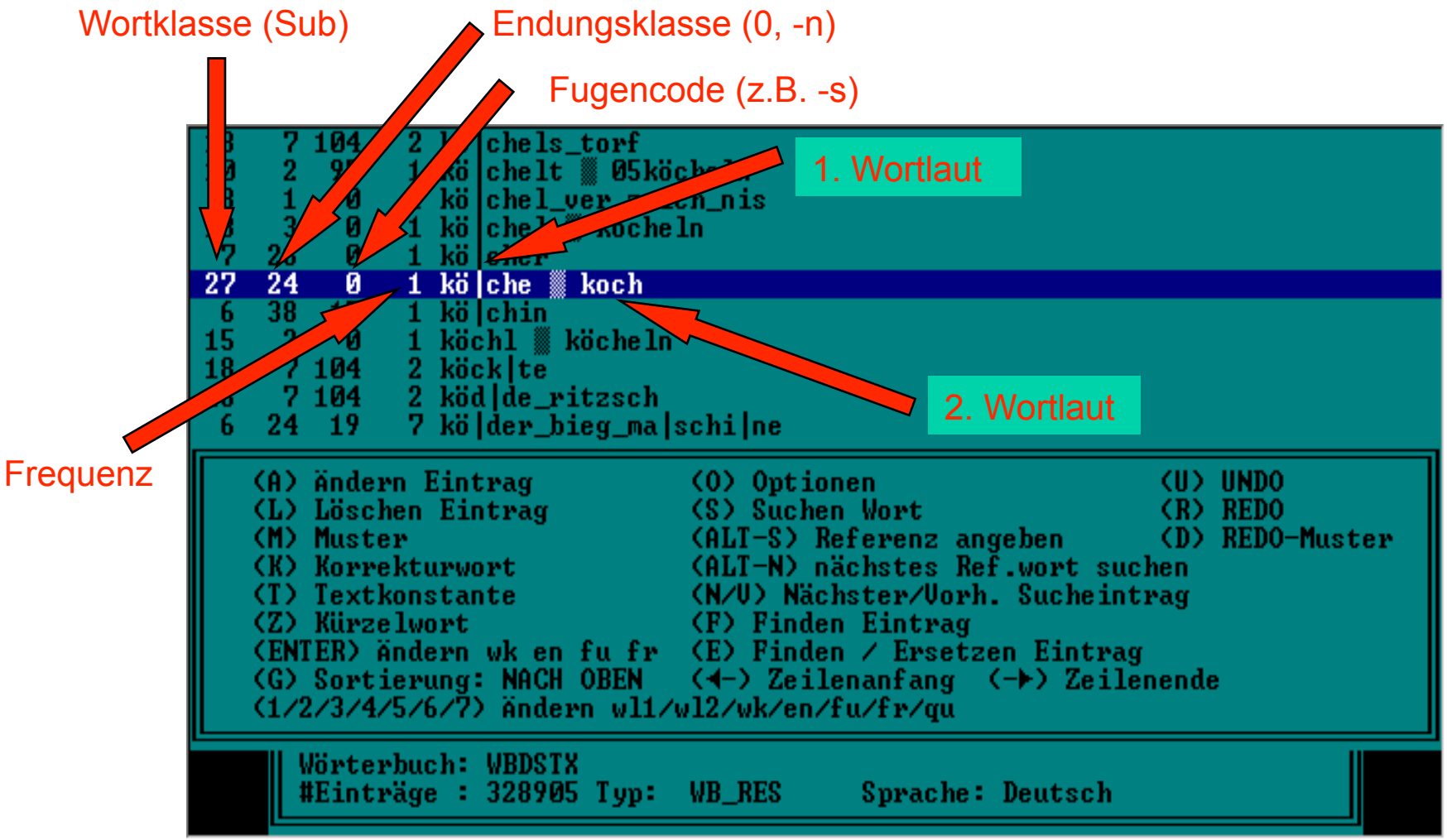
Interjektion

oh, au, ach

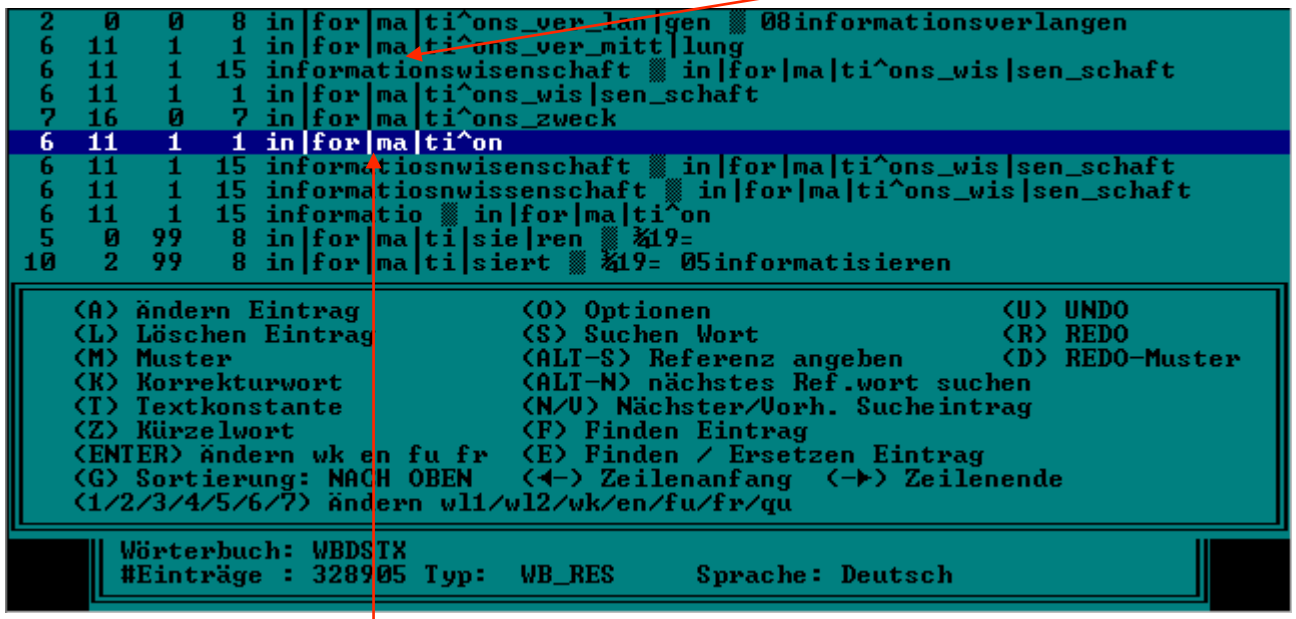
Morphy bei der Arbeit



Ein Grundformenlexikon

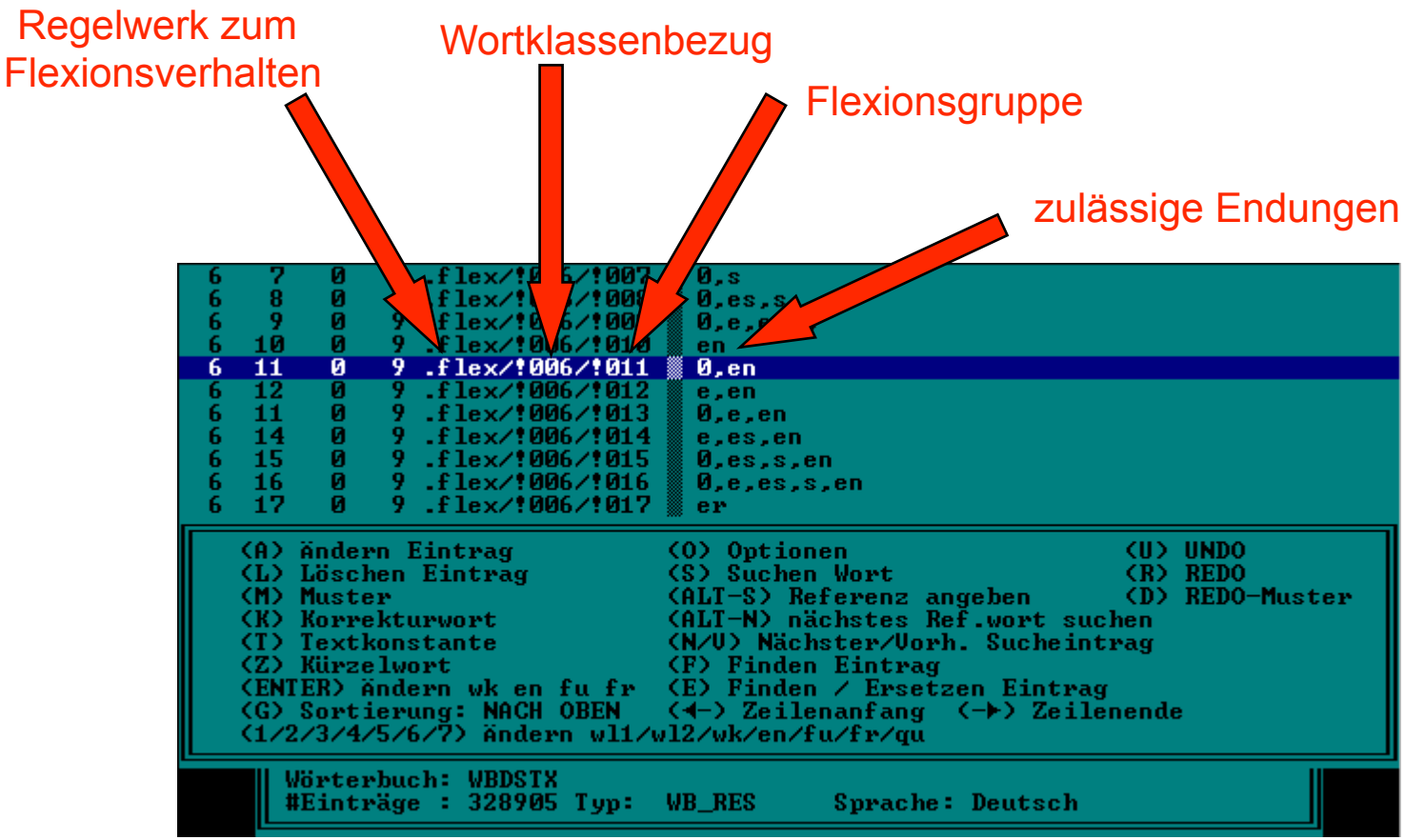


Beachte "Longest-Matching-Sortierung"



Eingabestring "Informationen" führt zu Lexikoneintrag "Information"

Grundform zu Wortform



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

Quelldaten

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

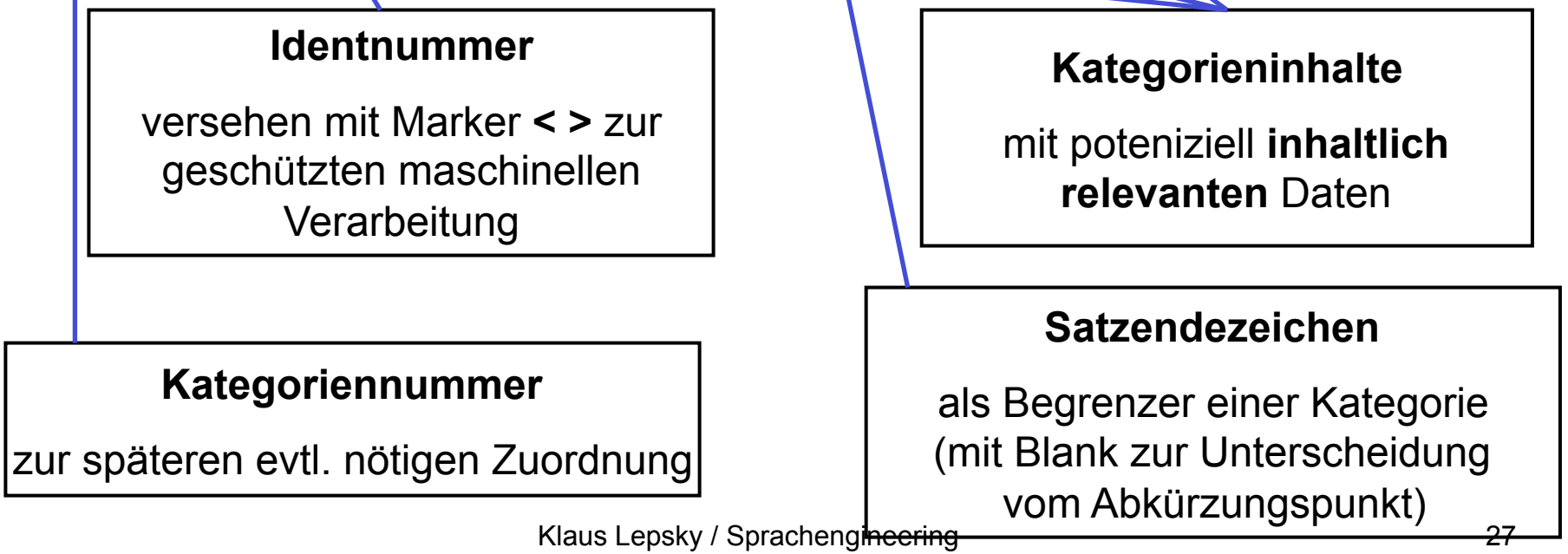
Umsetzung in das IDX-Eingangsformat

<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 .



Das Ergebnis

<00006 .>
*94 020 <0>
94 :
95 Aufbau <7>
96 und <1>
97 Pflege <6>
98 komplexer -> komplex <10>
99 natürlichsprachig <10>
100 basierter -> basiert <10>
100 basierter -> basieren <5>
101 Dokumentationssprachen ->
Dokumentationssprache <6>
102 (
102 Thesauri -> Thesaurus <7>
102)
103 .
*104 025 <0>
104 :
105 Aktuelle -> aktuell <10>
106 Tendenzen -> Tendenz <6>
107 und <1>
108 kritische -> kritisch <10>
109 Analyse <6>

Identnummer

Grundformen

110 einer -> ein <1>
110 einer -> ein <14>
111 ausgewählten -> ausgewählt <10>
111 ausgewählten -> auswählen <5>
112 autonomen -> autonom <10>
113 Thesaurus-Software <6>
114 für <1>
115 Personal -> personal <10>
115 Personal <8>
116 Computer <7>
117 (
117 PC <3>
117)
118 .
*119 100 <0>
119 :
120 INDEX -> Index <7>
121 .

Wortklasse

Zeichenkettenzähler

Grundformreduzierung mit IDX IV

Import in die Datenbank

00006* Analyse# Aufbau# Computer# Dokumentationssprache# Index# Personal# Pflege# Tendenz# Thesaurus# Thesaurus-Software# aktuell# ausgewählt# auswählen# autonom# basieren# basiert# komplex# kritisch# natürlichsprachig# personal

Speicherformat

(Komma delimited)

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 Informatio theoretischen Einführung eine INDEX 3.1
SPRACHE	d
OBJEKT	INDEX
Indexate	Analyse# Aufbau# Computer# Dokumentationssprache# Index# Personal# Pflege# Tendenz# Thesaurus# Thesaurus-Software# aktuell# ausgewählt# auswählen# autonom# basieren# basiert# komplex# kritisch# natürlichsprachig# personal

Import

in zusätzliche Kategorie

(kann für den Indexaufbau genutzt werden)

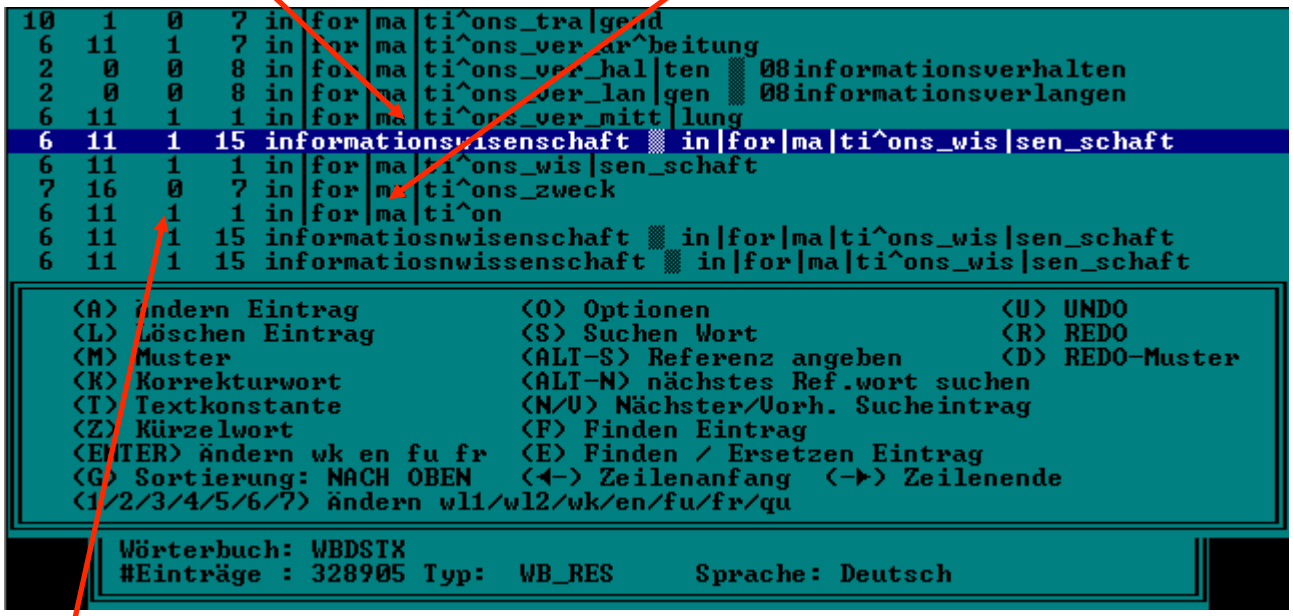
s

2.3 Kompositumerkennung I

Informationswirtschaft

Zerlegungsversuch über "Information"

nicht im Lexikon!



Fugencode 1 erlaubt Fugen-s (Regelwerk)

Suchstring "Information-s" ist identifiziert

Kompositumerkennung II

Fortsetzung der Identifizierung mit "Wirtschaft"
und Beenden der Kompositumanalyse

Aber Achtung!

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

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

"Baumangel"

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

Kompositumzerlegung mit IDX

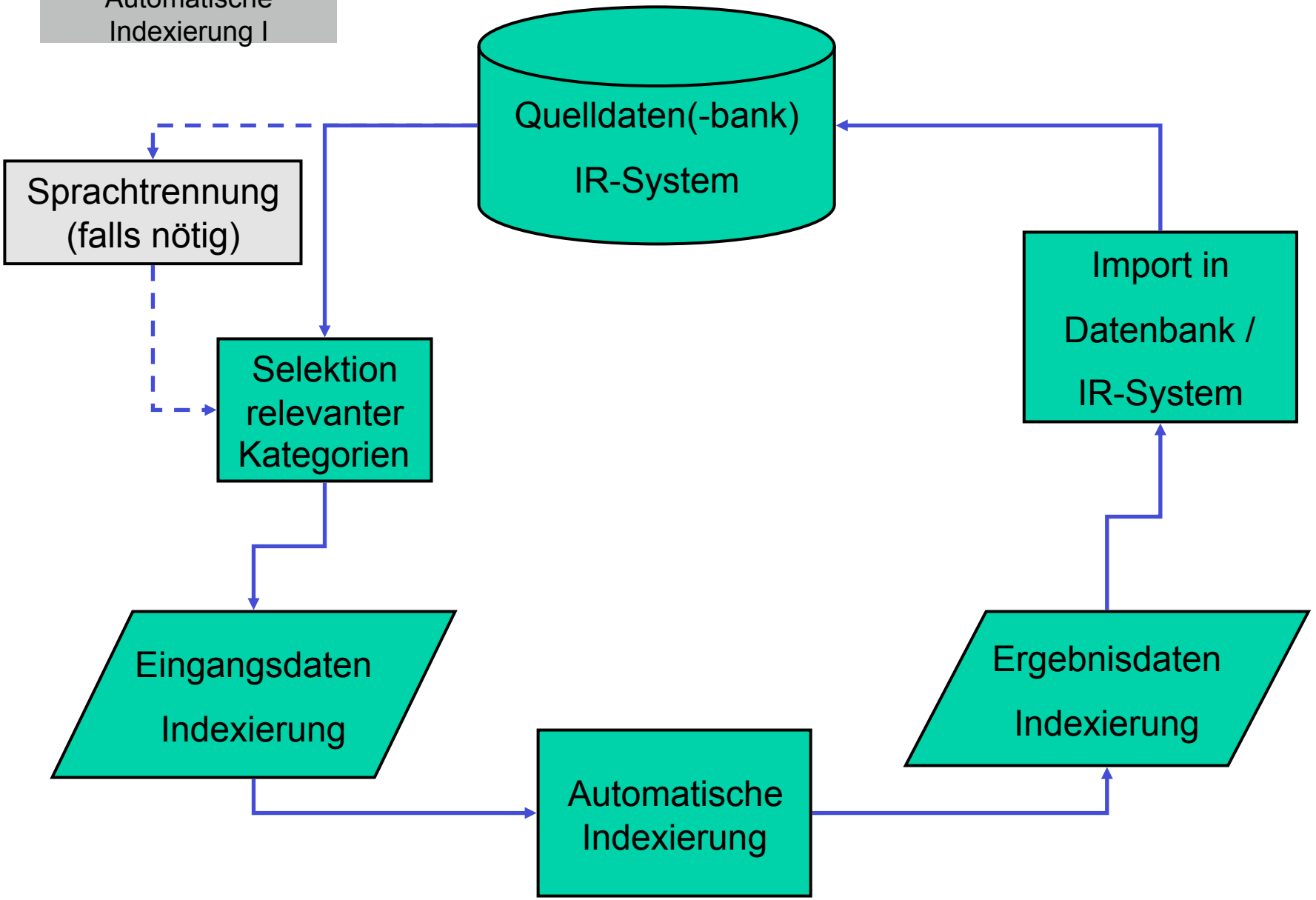
```
<00097 .>
*4638 020 <0>
4638 :
4639 a
4639 Der -> der <1>
4640 Einsatz <7>
4641 des <1>
4642 Automatischen -> automatisch <10>
<$B=4643,4645>
4643 Indexierungs- & Retrievalsystems -> Indexierungssystem <8> :0: System <8>
4643 Indexierungs- & Retrievalsystems -> Indexierungssystem <8> :3: Indexierung <6>
4643 Indexierungs -> Indexierung <6>
4643 -
4644 und <1>
4645 Retrievalsystems -> Retrievalsystem <8> :0: System <8>
4645 Retrievalsystems -> Retrievalsystem <8> :3: Retrieval <7>
4645 Retrievalsystems -> Retrievalsystem <8> :3: Retrieval <8>
4646 (
4646 AIR -> Air <8>
4646 )
4647 im <1>
4648 Fachinformationszentrum <8> :0: Zentrum <8>
4648 Fachinformationszentrum <8> :3: Fachinformation <8>
4649 Karlsruhe <18>
4650 .
```

Wortbindestrichtilgung

0 – letzter Wortbestandteil

3 – nicht letzter Wortbestandteil

Automatische Indexierung I



(1) Erzeugung von grammatikalischen Grundformen

Datenbanken	⇨	Datenbank
Häuser	⇨	Haus !
Handel	⇨	Hand ?
Händel	⇨	Hand ??

(2) Zerlegung von Komposita in sinnvolle Teilwörter

Informationswissenschaft	⇨	Information, Wissenschaft
Wissensdurst	⇨	Wissen, Durst
Wahnsinn	⇨	Wahn !, Sinn
Wirtschaft	⇨	Wirt, Schaft ?
Verbrechen	⇨	Verb, Rechen ??

(3) Bildung von Wortableitungen (Derivationen), z.B. Substantivierung von Adjektiven und Verben

wissenschaftlich	⇨	Wissenschaft
gefährlich	⇨	Gefahr
ging	⇨	Gang

(4) Erkennungsmechanismen für Mehrwortbegriffe

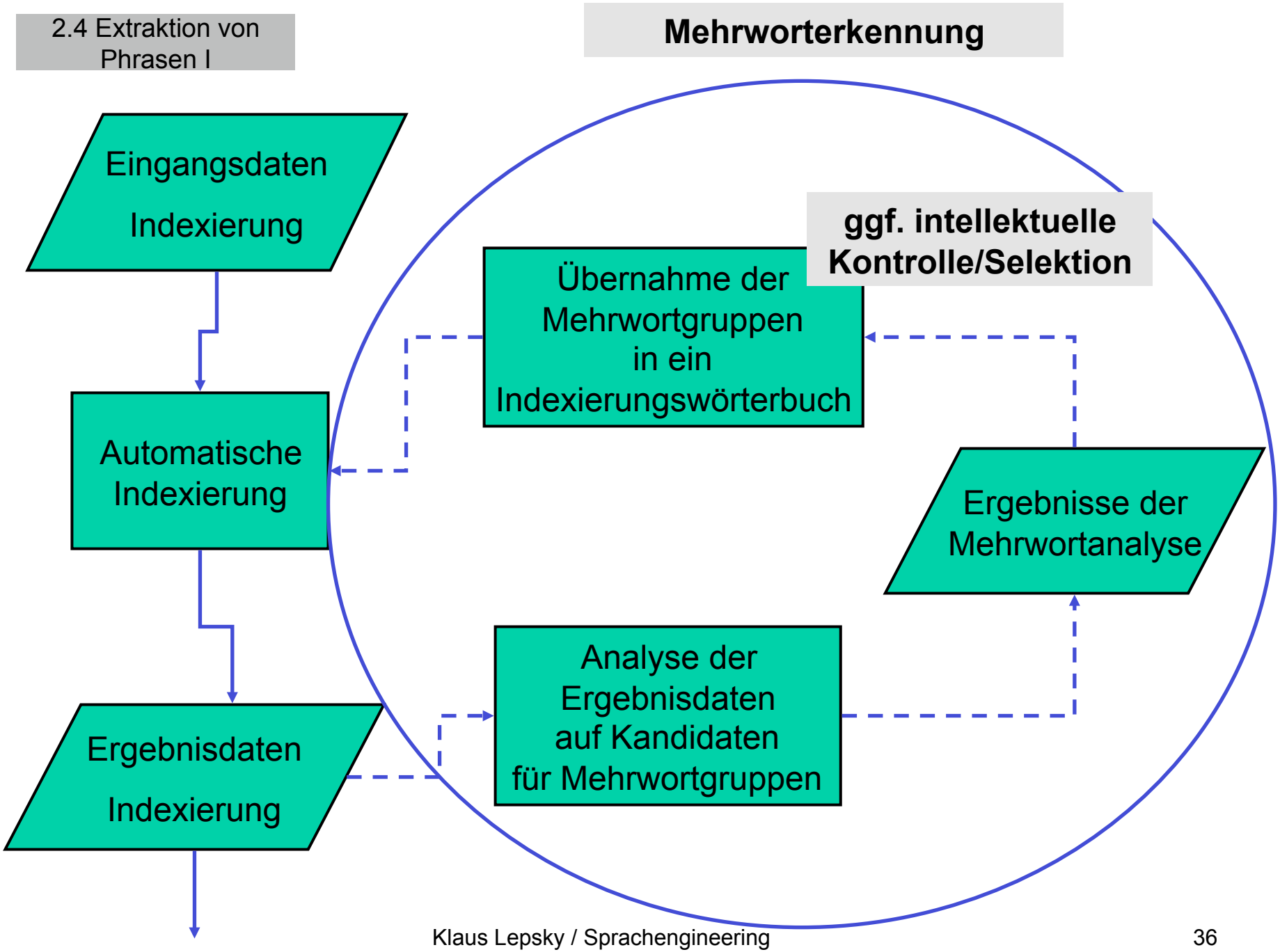
"Information und Dokumentation"

"juristische Person"

Johann Sebastian Bach	⇨	Bach, Johann Sebastian
... Bach ... Forelle	⇨	Bach (Gewässer)
... nahm ... teil	⇨	Teilnahme

(5) Möglichkeiten der begrifflichen Unterscheidung auf der Bedeutungsebene (Disambiguierung), z.B.

- zur Erkennung von Eigennamen (s.o.)
- zur Differenzierung von Homographen


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

Erkannte Kandidaten für Mehrwortbegriffe

32 5 0 6 0 0 2 automatische abhilfe in aussicht S5=AHA
24 3 0 6 0 0 2 automatische akquisition S3=A
23 3 0 6 0 0 2 automatische bestimmung S3=B
26 3 0 6 0 0 2 automatische deskribierung S3=D
33 3 0 6 0 0 2 automatische dokumenterschließung S3=D
35 3 0 6 0 0 2 automatische dokumentklassifikation S3=D
25 3 0 6 0 0 2 automatische erschließung S3=E
24 3 0 6 0 0 2 automatische gruppierung S3=G
24 3 0 6 0 0 2 automatische indexierung S3=I
41 5 0 6 0 0 2 automatische indexierung zur erschließung S5=IME
32 3 0 6 0 0 2 automatische inhaltserschließung S3=I
27 3 0 6 0 0 2 automatische klassifikation S3=K
21 3 0 6 0 0 2 automatische maschine S3=M
20 3 0 6 0 0 2 automatische methode S3=M
22 3 0 6 0 0 2 automatische recherche S3=R
34 3 0 6 0 0 2 automatische rechtschreibkorrektur S3=
22 3 0 6 0 0 2 automatische selektion S3=S
24 3 0 6 0 0 2 automatische verknüpfung S3=V
28 3 0 6 0 0 2 automatische vollindexierung S3=V
32 3 0 6 0 0 2 automatische wortformenreduktion S3=W
39 5 0 6 0 0 2 automatisierung in der sacherschließung S50AGS

alle Verbindungen

von

Adjektiv und **Substantiv**

Verbindungen

mit

Präpositionen

Einbeziehung von semantischen Beziehungen in die automatische Indexierung

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]

Indexierungsergebnis mit Thesaurusrelationen

```
<00001 .>
*1 020 <0>
 1 :
 2 a
 2 Die -> die <1>
 3 Aufgabenteilung <6> :0: Teilung <6> ## (1) Aufgabenteilung
 3 Aufgabenteilung <6> :3: Aufgabe <6> ## (1) Aufgabenteilung
 4 zwischen <1>
 5 Wortschatz <7> :0: Schatz <7> ## (1) Wortschatz
 5 Wortschatz <7> :1: Lexikon Linguistik <6> ## (1) Wortschatz
 5 Wortschatz <7> :1: Vokabular <8> ## (1) Wortschatz
 5 Wortschatz <7> :1: Terminologie Wortschatz <7> ## (1) Wortschatz
 5 Wortschatz <7> :1: Lexik <6> ## (1) Wortschatz
 5 Wortschatz <7> :3: Wort <8> ## (1) Wortschatz
 6 und <1>
 7 Grammatik <6>
 8 in <1>
 9 einer -> ein <1>
 9 einer -> ein <14>
 10 Indexsprache <6> :500: Sprache <6> ## (1) Indexsprache
 10 Indexsprache <6> :503: Index <7> ## (1) Indexsprache
 11 .
*
```

durch Thesauruseinträge erzeugte Relationierungen (Kennung 1)

Ergebnis: zusätzliche Sucheinstiege im semantischen Umfeld der Wortform!

Indexierungsergebnis mit vollständiger Funktionalität

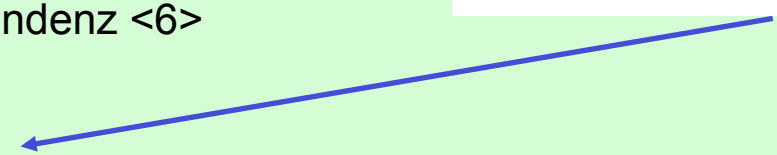
```
<00006 .>
*94 020 <0>
 94 :
 95 Aufbau <7> :4: aufbauen <5> ## (1) Aufbau
 96 und <1>
 97 Pflege <6>
 98 komplexer -> komplex <10>
 99 natürlichsprachig <10> :0: sprachig <10> ## (1) natürlichsprachig
 99 natürlichsprachig <10> :3: natürlich <10> ## (1) natürlichsprachig
 100 basierter -> basiert <10>
 100 basierter -> basieren <5>
 101 Dokumentationssprachen -> Dokumentationssprache <6> :0: Sprache <6> ## (1)
Dokumentationssprache
 101 Dokumentationssprachen -> Dokumentationssprache <6> :3: Dokumentation
<6> ## (1) Dokumentationssprache
 102 (
 102 Thesauri -> Thesaurus <7> :1: Deskriptor Verzeichnis <8> ## (1) Thesaurus
 102 Thesauri -> Thesaurus <7> :1: Deskriptorsprache <6> ## (1) Thesaurus
 102 )
 103 .
*104 025 <0>
```

Grundformen

zerlegte Komposita

```
*104 025 <0>
 104 :
 105 Aktuelle -> aktuell <10>
 106 Tendenzen -> Tendenz <6>
 107 und <1>
<$M2=108,109>
 108 kritische Analyse <6>
 108 kritische -> kritisch <10>
 109 Analyse <6> :1: Analytik <6> ## (1) Analyse
 109 Analyse <6> :3: Analysieren <8> ## (1) Analyse
 110 einer -> ein <1>
 110 einer -> ein <14>
 111 ausgewählten -> ausgewählt <10>
 111 ausgewählten -> auswählen <5> :0: wählen <5> ## (1) auswählen
 112 autonomen -> autonom <10>
 113 Thesaurus-Software <6> :500: Software <6> ## (1) Thesaurus-Software
 113 Thesaurus-Software <6> :503: Thesaurus <6> ## (1) Thesaurus-Software
```

Mehrwortbegriffe



```
114 für <1>
<$M1=115,116>
 115 Personal Computer <7>
 115 Personal -> personal <10> :4: Person <6> ## (1) personal
 115 Personal <8> :1: Mitarbeiter <7> ## (1) Personal
 116 Computer <7> :1: Computersystem Computer <7> ## (1) Computer
 116 Computer <7> :1: Rechner <7> ## (1) Computer
 116 Computer <7> :1: Rechenautomat <7> ## (1) Computer
 116 Computer <7> :1: Rechenanlage <6> ## (1) Computer
 116 Computer <7> :1: Elektronischer Rechenautomat <7> ## (1) Computer
 116 Computer <7> :1: Elektronische Rechenanlage <6> ## (1) Computer
 116 Computer <7> :1: Elektronenrechner <7> ## (1) Computer
 116 Computer <7> :1: DVA <74> ## (1) Computer
 116 Computer <7> :1: Digitalrechner <7> ## (1) Computer
 116 Computer <7> :1: Digitaler Computer <7> ## (1) Computer
 117 (
 117 PC <3>
 117 )
 118 .
*119 100 <0>
 119 :
 120 INDEX -> Index <7>
 121 .
```

Wortableitung

Thesaurusrelationen

Titelsatz und Indexierungsdaten

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
Indexate	00006* Analyse; Analysieren; Analytik; Aufbau; Computer; Computersystem Computer; DVA; Deskriptor Verzeichnis; Deskriptorsprache; Digitaler Computer; Digitalrechner; Dokumentation; Dokumentationssprache; Elektronenrechner; Elektronische Rechenanlage; Elektronischer Rechenautomat; Index; Mitarbeiter; PC; Person; Personal; Personal Computer; Pflege; Rechenanlage; Rechenautomat; Rechner; Software; Sprache; Tendenz; Thesaurus; Thesaurus-Software; aktuell; aufbauen; ausgewählt; auswählen; autonom; basieren; basiert; ein; für; komplex; kritisch; kritische Analyse; natürlich; natürlichsprachig; personal; sprachig; und; wählen

Indexierung einer dpa-Meldung

<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

dpa-Meldung II

```
<00001 .>
 1 Gesellschaft <6>
 1 :
 2 Strahlenrisiko <8> :1: Radiation hazard <1> ## (1) Strahlenrisiko
 2 Strahlenrisiko <8> :1: Strahlungsgefährdung <6> ## (1) Strahlenrisiko
 2 Strahlenrisiko <8> :1: Strahlengefährdung <6> ## (1) Strahlenrisiko
 2 Strahlenrisiko <8> :1: Strahlungsrisiko <8> ## (1) Strahlenrisiko
 2 Strahlenrisiko <8> :0: Risiko <8> ## (1) Strahlenrisiko
 2 Strahlenrisiko <8> :3: Strahl <7> ## (1) Strahlenrisiko
 3 wird -> werden <4>
 4 drastisch <10>
 5 unterschätzt -> unterschätzen <5> :0: schätzen <5> ## (1) unterschätzen
 6 =
 7 Bremen <18> :1: Hansestadt Bremen <18> ## (1) Bremen
 7 Bremen <18> :1: Bremen Land <7> ## (1) Bremen
 8 (
 8 dpa <1>
 8 )
 9 - <1>
<$M2=11,12>
 11 drastische Fehleinschätzung <6>
 11 drastische -> drastisch <10>
 12 Fehleinschätzung <6> :0: Einschätzung <6> ## (1) Fehleinschätzung
 13 des <1>
```

dpa-Meldung III

14 Strahlenrisikos -> Strahlenrisiko <8> :1: Radiation hazard <1> ## (1) Strahlenrisiko
14 Strahlenrisikos -> Strahlenrisiko <8> :1: Strahlungsgefährdung <6> ## (1) Strahlenrisiko
14 Strahlenrisikos -> Strahlenrisiko <8> :1: Strahlengefährdung <6> ## (1) Strahlenrisiko
14 Strahlenrisikos -> Strahlenrisiko <8> :1: Strahlungsrisiko <8> ## (1) Strahlenrisiko
14 Strahlenrisikos -> Strahlenrisiko <8> :0: Risiko <8> ## (1) Strahlenrisiko
14 Strahlenrisikos -> Strahlenrisiko <8> :3: Strahl <7> ## (1) Strahlenrisiko
15 hat -> haben <4>
16 die <1>
17 Gesellschaft <6>
18 für <1>
19 Strahlenschutz <7> :1: Strahlenschutzvorsorge <6> ## (1) Strahlenschutz
19 Strahlenschutz <7> :0: Schutz <7> ## (1) Strahlenschutz
19 Strahlenschutz <7> :3: Strahl <7> ## (1) Strahlenschutz
20 der <1>
21 Wirtschaft <6> :500: Schaft <7> ## (1) Wirtschaft
21 Wirtschaft <6> :1: Ökonomie Wirtschaft <6> ## (1) Wirtschaft
21 Wirtschaft <6> :1: Wirtschaftsleben <8> ## (1) Wirtschaft
21 ,
22 der <1>
23 Politik <6> :1: Politische Entwicklung <6> ## (1) Politik
23 Politik <6> :1: Politische Lage <1> ## (1) Politik
23 Politik <6> :1: Staatspolitik <6> ## (1) Politik
24 und <1>
25 einer -> ein <1>
25 einer -> ein <14>
26 `
26 industriefreundlichen -> industriefreundlich <10>

dpa-Meldung IV

```
27 Wissenschaft <6> :500: Schaft <7> ## (1) Wissenschaft
27 Wissenschaft <6> :1: Wissenschaften <8> ## (1) Wissenschaft
27 Wissenschaft <6> :1: Bürgerliche Wissenschaft <6> ## (1) Wissenschaft
27 "
28 vorgeworfen -> vorwerfen <5> :0: werfen <4> ## (1) vorwerfen
28 .
29 Dies -> dies <1>
30 habe -> haben <5>
31 dazu <1>
32 beigetragen <10>
32 beigetragen -> beitragen <5> :4: Beitrag <7> ## (1) beitragen
32 beigetragen -> beitragen <5> :4: Beiträger <7> ## (1) beitragen
32 ,
33 dass <1>
34 es <1>
35 in <1>
36 Deutschland <18>
37 heute -> heuen <5>
37 heute <1>
38 mehr <1>
38 mehr <30>
39 als <1>
40 30 <Z>
41 000 <Z>
42 anerkannte -> anerkennen <4>
43 Fälle -> Fall <7>
44 von <1>
```

dpa-Meldung V

45 Berufskrankheiten -> Berufskrankheit <6> :1: Arbeitsbedingte Krankheit <6> ## (1) Berufskrankheit
45 Berufskrankheiten -> Berufskrankheit <6> :0: Krankheit <6> ## (1) Berufskrankheit
45 Berufskrankheiten -> Berufskrankheit <6> :3: Beruf <7> ## (1) Berufskrankheit
46 gebe -> geben <4>
46 ,
47 die <1>
48 durch <1>
49 Arbeiten -> arbeiten <5>
49 Arbeiten -> Arbeit <6> :1: Erwerbsarbeit <6> ## (1) Arbeit
49 Arbeiten -> Arbeit <6> :4: Arbeiten <68> ## (1) Arbeit
50 im <1>
51 Bereich <7> :1: Feld Philosophie <6> ## (1) Bereich
52 der <1>
53 Atomindustrie <6> :1: Kerntechnische Industrie <6> ## (1) Atomindustrie
54 unter <10>
54 unter <1>
55 mangelhaften -> mangelhaft <10> :4: mangeln <5> ## (1) mangelhaft
55 mangelhaften -> mangelhaft <10> :3: Mangel <7> ## (1) mangelhaft
56 Schutzbestimmungen -> Schutzbestimmung <6> :0: Bestimmung <6> ## (1) Schutzbestimmung
56 Schutzbestimmungen -> Schutzbestimmung <6> :3: Schutz <7> ## (1) Schutzbestimmung
57 hervor <1>
58 gerufen <10> :4: rufen <4> ## (1) gerufen
58 gerufen <10> :4: Ruf <7> ## (1) gerufen
59 worden -> werden <5>
60 seien -> sein <4>

dpa-Meldung VI

60 ,
61 kritisierte -> kritisieren <5>
61 kritisierte -> kritisiert <10>
62 der <1>
63 Präsident <7>
64 der <1>
65 Gesellschaft <6>
65 ,
66 Sebastian <17>
67 Pflugbeil <8>
68 (
68 Berlin <18> :1: Großberlin <18> ## (1) Berlin
68).
69 Die -> die <1>
70 Strahlenschutzverordnung <6> :0: Verordnung <6> ## (1) Strahlenschutzverordnung
70 Strahlenschutzverordnung <6> :3: Strahlenschutz <7> ## (1) Strahlenschutzverordnung
71 des <1>
72 Bundes -> Bunde <18>
73 habe -> haben <5>
74 seit <1>
75 Jahrzehnten -> Jahrzehnt <8>
76 ein <1>
77 unterschätztes -> unterschätzen <5> :0: schätzen <5> ## (1) unterschätzen
78 Risiko <8>
79 zur -> zu <1>
80 Grundlage <6>
80 ,
81 sagte -> sagen <5>

dpa-Meldung VII

```
82 Pflugbeil <8>
83 am <3>
84 Donnerstag <7>
85 in <1>
86 Bremen <18> :1: Hansestadt Bremen <18> ## (1) Bremen
86 Bremen <18> :1: Bremen Land <7> ## (1) Bremen
86 .
87 Dort -> dort <1>
88 beginnt -> beginnen <4>
89 am <3>
90 Freitag <7> :0: Tag <7> ## (1) Freitag
90 Freitag <7> :3: frei <10> ## (1) Freitag
91 der <1>
92 zweitägige -> zweitägig <10> :0: tägig <10> ## (1) zweitägig
92 zweitägige -> zweitägig <10> :3: zwei <14> ## (1) zweitägig
93 internationale -> international <10>
94 Kongress <7>
95 `
95 Strahlenschutz <7> :1: Strahlenschutzvorsorge <6> ## (1) Strahlenschutz
95 Strahlenschutz <7> :0: Schutz <7> ## (1) Strahlenschutz
95 Strahlenschutz <7> :3: Strahl <7> ## (1) Strahlenschutz
96 nach <1>
97 der <1>
98 Jahrtausendwende <6> :0: Wende <6> ## (1) Jahrtausendwende
98 Jahrtausendwende <6> :1: Jahrtausendende <8> ## (1) Jahrtausendwende
98 Jahrtausendwende <6> :1: Jahrtausendwechsel <7> ## (1) Jahrtausendwende
98 Jahrtausendwende <6> :3: Jahrtausend <8> ## (1) Jahrtausendwende
98 ".
```

dpa-Meldung VIII

```
 99 Er <3>
 100 forderte -> fordern <5>
 101 eine -> einen <5>
 101 eine -> ein <14>
<$M2=102,103>
 102 deutliche Senkung <6>
 102 deutliche -> deutlich <10> :4: deuten <5> ## (1) deutlich
 103 Senkung <6>
 104 des <1>
 105 Grenzwertes -> Grenzwert <7> :1: Zulässiger Grenzwert <7> ## (1) Grenzwert
 105 Grenzwertes -> Grenzwert <7> :0: Wert <7> ## (1) Grenzwert
 105 Grenzwertes -> Grenzwert <7> :4: Grenze <6> ## (1) Grenzwert
 105 Grenzwertes -> Grenzwert <7> :3: Grenze <6> ## (1) Grenzwert
 106 für <1>
 107 beruflich <10>
 108 von <1>
 109 Strahlen -> strahlen <5>
 109 Strahlen -> Strahl <7>
<$M2=110,111>
 110 betroffene Personen -> betroffene Person <6>
 110 betroffene -> betroffen <10>
 111 Personen -> Person <6>
 111 .

                    IDX, VER. 28/05/2002 (C) SOFTEX
```

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

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

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

6. Interpretation der Ergebnisse

- Was ist gut?
- Warum?

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)

„Automatic Abstracting“ meint die automatische Erzeugung sinnvoller Zusammenfassungen von Texten

Experte: Tarifstreit gefährdet Arbeitsplätze in Ostdeutschland =

Hamburg (dpa) - Ein zu hoher Tarifabschluss im öffentlichen Dienst gefährdet nach Ansicht von Experten und öffentlichen Arbeitgebern massiv Arbeitsplätze in Ostdeutschland. `Der ÖTV-Streik käme zur Unzeit", sagte der Präsident des Instituts für Wirtschaftsforschung in Halle, Rüdiger Pohl, der Zeitung `Welt am Sonntag". Der Streik laufe auf einen Konflikt hinaus, der mit Enttäuschungen enden müsse.

Wenn die Gewerkschaften mehr als eine zusätzliche Steigerung des Lohnes um 0,2 Prozentpunkte wollten, könnten die ostdeutschen Länder dies nicht finanzieren, sagte der Wirtschaftsexperte Pohl. `Das Geld ist nicht da. Das ist die simple Wahrheit." In den Kommunen und Landesverwaltungen werde es verstärkt zu betriebsbedingten Kündigungen kommen, die bisher vermieden worden seien.

Auch der Verhandlungsführer der Länder, Sachsens Finanzminister Georg Milbradt (CDU), hält betriebsbedingte Kündigungen für möglich. Der Schlichterspruch bedeute de facto im Osten eine Lohnerhöhung von acht Prozent, sagte Mildbradt dem Nachrichtenmagazin `Der Spiegel". `Das kann nur mit einem radikalem Stellenabbau im öffentlichen Dienst kompensiert werden.„

Nach der Ablehnung des Schlichterspruchs beginnen an diesem Montag die Urabstimmungen. Wenn die Mitglieder für einen Arbeitskampf stimmen, wäre mit ersten Streiks nach Pfingsten zu rechnen. Nach dem Schiedsspruch sollten die Einkommen rückwirkend zum 1. April um 1,8 Prozent sowie ein Jahr später um weitere 2,2 Prozent erhöht werden. Die Ostgehälter sollten bis 2002 in Stufen von derzeit 86,5 Prozent auf 90 Prozent des Westniveaus steigen.

Ein Ergebnis

Copernic Summarizer – 50 % Reduktion

- Hamburg (dpa) - Ein zu hoher Tarifabschluss im öffentlichen Dienst gefährdet nach Ansicht von Experten und öffentlichen Arbeitgebern massiv Arbeitsplätze in Ostdeutschland.

- `Der ÖTV-Streik käme zur Unzeit", sagte der Präsident des Instituts für Wirtschaftsforschung in Halle, Rüdiger Pohl, der Zeitung `Welt am Sonntag".

- Der Streik laufe auf einen Konflikt hinaus, der mit Enttäuschungen enden müsse.

- Wenn die Gewerkschaften mehr als eine zusätzliche Steigerung des Lohnes um 0,2 Prozentpunkte wollten, könnten die ostdeutschen Länder dies nicht finanzieren, sagte der Wirtschaftsexperte Pohl.

- In den Kommunen und Landesverwaltungen werde es verstärkt zu betriebsbedingten Kündigungen kommen, die bisher vermieden worden seien.

- Georg Milbradt (CDU), hält betriebsbedingte Kündigungen für möglich.

- Der Schlichterspruch bedeute de facto im Osten eine Lohnerhöhung von acht Prozent, sagte Mildbradt dem Nachrichtenmagazin `Der Spiegel".

- `Das kann nur mit einem radikalem Stellenabbau im öffentlichen Dienst kompensiert werden."

- Nach der Ablehnung des Schlichterspruchs beginnen an diesem Montag die Urabstimmungen.

Ein zweites Ergebnis

Copernic Summarizer – 75% Reduzierung

- Hamburg (dpa) - Ein zu hoher Tarifabschluss im öffentlichen Dienst gefährdet nach Ansicht von Experten und öffentlichen Arbeitgebern massiv Arbeitsplätze in Ostdeutschland.

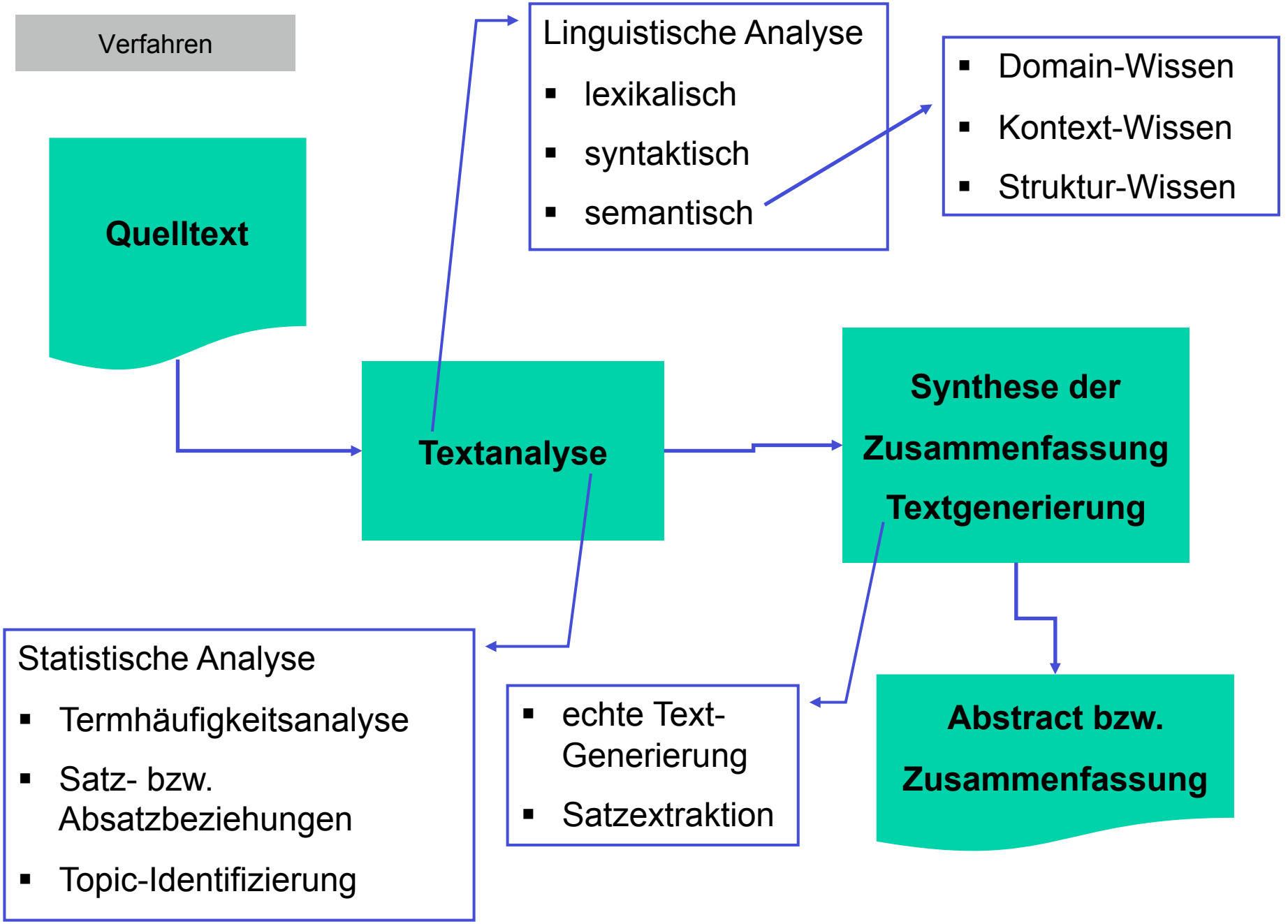
- In den Kommunen und Landesverwaltungen werde es verstärkt zu betriebsbedingten Kündigungen kommen, die bisher vermieden worden seien.

- Der Schlichterspruch bedeute de facto im Osten eine Lohnerhöhung von acht Prozent, sagte Mildbradt dem Nachrichtenmagazin `Der Spiegel".

- `Das kann nur mit einem radikalem Stellenabbau im öffentlichen Dienst kompensiert werden."

Copernic Summarizer – 90% Reduzierung

- Hamburg (dpa) - Ein zu hoher Tarifabschluss im öffentlichen Dienst gefährdet nach Ansicht von Experten und öffentlichen Arbeitgebern massiv Arbeitsplätze in Ostdeutschland.


Verfahren
Quelltext
Linguistische Analyse
lexikalisch
syntaktisch
semantisch
Domain-Wissen
Kontext-Wissen
Struktur-Wissen
Textanalyse
Synthese der Zusammenfassung Textgenerierung
Statistische Analyse
Termhäufigkeitsanalyse
Satz- bzw. Absatzbeziehungen
Topic-Identifizierung
echte Text-Generierung
Satzextraktion
Abstract bzw. Zusammenfassung

Die Satzglieder

- **Subjekt**

bezeichnet den Satzgegenstand und ist Substantiv/Nomen

Die **Katze** läuft.
Es schneit.

- **Prädikat**

beschreibt Handlungen und ist Verb

Die Katze **läuft**.
Es **schneit**.

- **Objekt**

bezeichnet das Ziel bzw. Ergebnis einer Handlung

Peter lernt gerne **Grammatik**. (Akkusativobjekt)
Die Schüler geben **dem Lehrer** immer richtige Antworten. (Dativobjekt)
Sie bezichtigt ihn **der Lüge**. (Genitivobjekt)
Er hängt das Bild **an die Wand**. (Präpositionalobjekt)

Satz mich einem Syntax korrekte macht zu.

Syntax

Korrekte Syntax macht mich zu einem Satz.

Syntax

beschäftigt sich damit, wie Wörter zu **korrekten** Sätzen zusammengefügt werden;

legt die strukturelle Rolle der Wörter in Sätzen fest;

Diese Bedeutung ist ohne Satz sinnlos.

Semantik

Dieser Satz ist ohne Bedeutung sinnlos.

Semantik

beschäftigt sich damit, welche **Bedeutung** Wörter haben (Wortsemantik);

wie die **kombinierten Bedeutungen** mehrerer Wörter in Sätzen zu Satzbedeutungen werden (Satzsemantik);

Aber: Der Kontext der Verwendung des Satzes wird nicht untersucht (kontext-unabhängige Bedeutung).

Pragmatik

Verschwinde!

Wohin?

Was?

Soeben war es noch an seinem Platz.

Wer?

Warum?

Wann genau?

Wo?

Hast du Hunger? – Ja bitte.

Häh?

Pragmatik

beschäftigt sich damit, welche (unterschiedlichen) Bedeutungen Sätze in Abhängigkeit von der Situation haben, in der sie verwendet werden.

untersucht insbesondere die funktionale Bedeutung sprachlicher Ausdrücke

Übung 2

Analysieren Sie die folgenden Aussagen im Hinblick auf Syntax und Semantik:

(1) Sprache ist einer der fundamentalen Aspekte menschlichen Verhaltens und ist damit entscheidender Bestandteil unseres Lebens.

(2) Grüne Frösche haben große Nasen.

(3) Grüne Ideen haben große Nasen.

(4) Großes haben grüne Ideen Nasen.

(5)
```
x=0;
    For x < 10 Do;
        Print "Immer noch nicht";
        x=x+1;
    Print "Endlich!";
```

Elementare Funktionen der Sprache als Mittel der Kommunikation (nach Karl Bühler):

Darstellen, d.h. für Etwas stehen, Etwas symbolisieren.

Das Planetensystem hat als Mittelpunkt die Sonne, einen Fixstern von eher durchschnittlicher Größe, um den sich die Planeten in leicht elipsoiden Bahnen drehen.

Ausdruck in Abhängigkeit vom Sprecher (Symptom)

Heute fühl ich mich gar nicht gut.

Appell an einen Angesprochenen (Signal)

Vielleicht fährst du lieber langsam, denn es fängt an zu regnen.

Ziele automatischer Sprach- bzw. Textanalyse

1. Analyse von Sprache oder Texten

im Sinne des "Verstehens" von Sprache

Quellen sind geschriebene oder gesprochene natürliche und nicht-natürliche Sprache

Ziele sind:

Befehlseingabe, z.B. Bedienung von Programmen

Umwandlung von Quell-Sprache (natürlich und nicht-natürlich) in Systemsprache, z.B. im Information Retrieval oder in der Programmierung

Verarbeitung natürlicher Sprache, z.B.

Rechtschreibkontrolle, Diktiersysteme, Automatische Übersetzung, Automatische Indexierung, Abstracting

2. Generierung von Sprache bzw. Text

im Sinne des "Erzeugens" von Sprache für z.B.

Mensch-Maschine-Schnittstellen

Automatische Übersetzung, Abstracting

Robotersysteme? ELIZA?

Sprachanalyse lässt sich unterscheiden in **Satzanalyse** und Diskurs- bzw. Dialoganalyse

Satzanalyse umfasst wiederum **Syntaxanalyse** und semantische Analyse

Werkzeuge zur Syntaxanalyse heißen **Parser**, der Prozess der Syntaxanalyse heißt **Parsing**

Parser benötigen zur Syntaxanalyse **formale Grammatiken** und **Lexika**

Formale Grammatiken

dienen der **Beschreibung von Sprachen**, z.B. von formalen Sprachen (Programmiersprachen) und von natürlicher Sprache

sind geeignet, die Syntax formaler und natürlicher Sprachen zu analysieren

sind damit Basis für alle Aufgabenbereiche der Computerlinguistik

Phrasenstrukturgrammtiken

sind formale Grammatiken, die ein Set von Regeln umfassen, das grammatikalisch korrekte Sätze einer Sprache erzeugen kann, z.B. Sätze der natürlichen Sprache:

"Ich ging einkaufen."

"Rice flies like sand."

Phrasenstrukturgrammatiken (PSG) bestehen aus vier Komponenten

Terminale Symbole "T"

die Gesamtheit aller terminalen Symbole entspricht dem Vokabular, den Wörtern (oder Symbolen) einer Sprache

Haus

Baumstamm

der

alte

unmöglich

Einstein

Non-Terminale Symbole "N"

bilden die Menge aller grammatikalischen Symbole bzw. Strukturen, die in der Regel aus mehreren teminalen und nonterminalen Symbolen bestehen

NP (Nominalphrase),
z.B. alter Mann

VP (Verbalphrase),
z.B. ging schnell

PP (Präpositionalphrase),
z.B. an der Wand

Set von Regeln "P"

die Gesamtheit der Regeln, die für die Bildung von grammatikalischen Strukturen (Satzglieder, Sätze) zur Verfügung stehen

Regeln haben die allgemeine Form:

a -> b

wobei **a** eine Folge mehrerer Symbole aus T und N ist und **b** eine Folge von keinem oder mehreren Symbolen aus T und N ist:

NP -> ADJ N

Startsymbol "S"

einem Element aus der Menge der nonterminalen Symbole N

Das **Startsymbol S** definiert den Anfang des Satzbauprozesses, z.B. als:

S -> NP VP

S legt also fest, dass ein gültiger Satz aus einer Nominalphrase und einer Verbalphrase besteht

Auf der Basis der vier Elemente **Terminale, Non-Terminale, Regelset, Startsymbol** lassen sich gültige Sätze generieren:

N = {S}

T = {a,b,c}

P = {S -> aSc, S -> b}

erzeugt z.B.

S => aSc => abc

d.h. abc ist ein gültiger Satz der durch die Regeln P definierten Sprache

ebenso

S => aSc => aaScc => aabcc

und

S => aSc => aaScc => aaaSccc => aaabccc

oder allgemein

... => aaa...b...ccc

Alle Sätze, die aus einem Regelsystem abgeleitet werden können, bilden die durch diese Grammatik definierte Sprache.

Sprachgenerierung

Lexikon -> Regelsystem -> Sätze

Ein Programm, das die Ableitungen eines Satzes (in Bezug auf eine Grammatik) analysiert, ist ein **Parser**.

Sprachanalyse

Satz -> Regelsystem -> Lexikon

Alle Methoden des Sprachengineerung benötigen Verfahren zur Sprachgenerierung und/oder Sprachanalyse.

Übung 3

1. Welche Satztypen werden von folgender Grammatik erzeugt:

   S -> aA
   A -> bB
   B -> cA
   B -> d

2. Mit welcher Grammatik lassen sich folgende *Sätze* erzeugen:

   x,(x),((x)),(((x))),((((x)))), ...

3. Kennzeichnen Sie die Unterschiede zwischen Grammatik 1 und 2

4. Welche Sätze erzeugt

   S -> [S]
   S -> a

Reguläre PS-Grammatiken

umfassen nur Regeln der Formen

A -> b, A -> bC (rechtslinear)

A -> b, A -> Cb (linkslinear)

d.h., die linke Seite besteht immer aus einem Non-Terminal, die rechte entweder aus einem Terminal oder einem Terminal gefolgt von einem Non-Terminal (und umgekehrt für die linkslineare Form)

Kontextfreie PS-Grammatiken

haben die allgemeine Form

sog. Backus-Naur-Form

A -> x bzw. <A> ::= x

wobei A ein non-terminales Symbol ist und x eine Folge von keinem oder mehreren terminalen und non-terminalen Symbolen ist.

Beispiel einer kontextfreien Grammatik für natürliche Sprache

Benötigte Wortklassen

S	Substantiv
V	Verb
P	Präposition
A	Artikel
ADJ	Adjektiv
ADV	Adverb

Realisierung einer Wortklassendefinition als Regelsystem

<S> ::= Baum | Haus | Mann | Sommer
<V> ::= sein | ist | geht | ging | steht | gehen
<P> ::= in | über | an | im
<A> ::= der | ein
<ADJ> ::= alte | helle | buntes
<ADV> ::= langsam | gemächlich

Beispiel II

Alternative Möglichkeit der Wortklassendefinition durch **Lexikon** mit **Präterminalen**

L = {alte [ADJ], an [P], Baum [S], buntes [ADJ], der [A], ein [A], gehen [P], geht [V], gemächlich [ADV], ging [V], Haus [S], helle [ADJ], im [P], in [P], ist [V], langsam [ADV], Mann [S], sein [V], Sommer [S], steht [V], über [P], ...}

Vorteil:
Flexibilität durch Vokabularveränderung bzw. –erweiterung außerhalb des Regelwerks

Aufbau der Grammatik

Grammatik G = {PT, N, P, S}

Präterminale PT = {S, V, ADJ, ...}

Nonterminale N = {S, VP, NP, ...}

Regelwerk P

Startsymbol S

Lexikon L = {...}

Aufbau des Regelwerks

<Satz> ::= <Aussage> | <Frage> | <Befehl>

ein Satz ist entweder eine Aussage, eine Frage, oder ein Befehl

<Aussage> ::= <S> <V> | <S> <V> <S>

eine Aussage besteht aus einem Substantiv gefolgt von einem Verb, oder aus einem Substantiv gefolgt von einem Verb gefolgt von einem Substantiv:

Franz schläft. Franz isst Käse.

Einführung von Erweiterungsmöglichkeiten, um komplexere Sätze bilden zu können; Erweiterungen Links und Rechts werden allgemein definiert als:

<LxR> ::= <Lx> <x> <Rx>

Nonterminal <LxR> für die Wortklasse x besteht aus der linken Erweiterung <Lx> von x, gefolgt von x, gefolgt von der rechten Erweiterung <Rx> von x

Beispiel Substantiv

<LSR> ::= <LS> <S> <RS>

erlaubt linke und rechte Erweiterungen für Substantive

<LS> ::= <APOS> <ADJPOS>

lässt für linke Substantiv-Erweiterungen Artikelgruppen und Adjektivgruppen zu

<APOS> ::= <A> | null

definiert eine Artikelgruppe als Artikel oder leer

<ADJPOS> ::= <ADJ> | null

definiert eine Adjektivgruppe als Adjektiv oder leer

<RS> ::= <PS> | null

erlaubt rechte Erweiterungen als Präpositionalphrasen

Beispiel V

<PS> ::= <P> <S>

definiert Präpositionalphrasen als String aus Präposition und Substantiv

Beispiel

<LSR> ::= <A> <ADJ> <S> <P> <S>

die fleißige Studentin aus Köln

die Studentin

Studentin

Einführung von Satzergänzungen

<SE> ::= <ADV> | <PS> | null

Satzergänzungen bestehen aus Adverben oder Präpositionalphrasen oder sind leer

Beispiel VI

Einfügen von <LSR> und <SE> in die Definition für Aussagen

<Aussage> ::= <SE> <LSR> <SE> <LVR> <SE> |
<SE> <LSR> <SE> <LVR> <SE> <LSR> <SE>

Aussagen bestehen aus
Satzergänzung, Substantiv mit möglicher linker oder rechter Erweiterung, Satzergänzung, Verb mit möglicher linker oder rechter Erweiterung, Satzergänzung

oder
Satzergänzung, Substantiv mit möglicher linker oder rechter Erweiterung, Satzergänzung, Verb mit möglicher linker oder rechter Erweiterung, Satzergänzung, Substantiv mit möglicher linker oder rechter Erweiterung, Satzergänzung

Beispiel VII

Einführung des Symbols *Objekt*

<Aussage> ::= <SE> <LSR> <SE> <LVR> <SE> <Objekt> <SE>

die abhängige Substantivgruppe wird definiert als Objekt der Aussage/Handlung

<Objekt> ::= <LSR> | null

Objekt ist definiert als Substantiv mit möglicher linker und/ oder rechter Erweiterung oder leer

analoge Einführung der Symbole *Subjekt* und *Prädikat*

<Aussage> ::= <SE> <Subjekt> <SE> <Prädikat> <SE> <Objekt> <SE>

Aussagen bestehen aus Satzergänzung, Subjekt, Satzergänzung, Prädikat, Satzergänzung, Objekt, Satzergänzung

Beispiel VIII

Definition Subjekt

<Subjekt> ::= <LSR>

Subjekt besteht aus Substantiv mit möglicher linker und rechter Erweiterung

Definition Prädikat

<Prädikat> ::= <LVR>

Prädikat besteht aus Verb mit möglicher linker und rechter Erweiterung

Definition Objekt

<Objekt> ::= <LSR> | null

Objekt besteht aus Substantiv mit möglicher linker und rechter Erweiterung oder ist leer

Zusammenführung aller Elemente der kontextfreien PS-Grammatik für natürliche Sprache

<Satz> ::= <Aussage>

<Aussage> ::= <SE> <Subjekt> <SE> <Prädikat> <SE> <Objekt> <SE>

<SE> ::= <ADV> | <PS> | null

<PS> ::= <P> <S>

<Subjekt> ::= <LSR>

<LSR> ::= <LS> <S> <RS>

<LS> ::= <APOS> <ADJPOS>

<APOS> ::= <A> | null

<ADJPOS> ::= <ADJ> | null

<RS> ::= <PS> | null

Beispiel X

<Prädikat> ::= <LVR>

<LVR> ::= <LV> <V> <RV>

<LV> ::= <ADV> | null

<RV> ::= <ADV> | <PS> | null

<Objekt> ::= <LSR> | null

Übung 4

1. Bilden Sie gültige Sätze unter Verwendung des zu Grunde gelegten Regelwerks und Lexikons

2. Entscheiden Sie, ob folgende Sätze gültige Sätze sind:

   Der alte Mann geht langsam in ein buntes Haus.

   Der buntes Mann ging im Haus gemächlich über Sommer.

   Gemächlich steht sein Baum über Mann.

3. Welche Probleme werden von der Grammatik nicht behandelt?

4. Wie müsste die Grammatik verändert werden, um diese Probleme zu lösen?

Die Schwächen kontextfreier Grammatiken lassen sich durch die Vereinbarung von Bedingungen und Einschränkungen beheben, sog. **constraints**, z.B:

Harmonisierung von Plural und Singular zwischen Verb und Substantiv

Bäume steht

abzählbare Substantive verlangen den Artikel

Katze frisst

Harmonisierung zwischen Verb und Objekt

Die Katze frisst Mäuse

Die Katze schläft Mäuse

Weitere Typen von Grammatiken

kontextsensitive Grammatiken

bestehen aus Regeln der Form

x -> y

wobei **y** gleich viel oder mehr Symbole umfasst als **x**

beschreiben **rekursive Sprachen**, d.h. es lässt sich ein Programm schreiben, das entscheidet, ob ein gegebener Satz Element der Sprache ist oder nicht

sind mächtiger als kontextfreie Grammatiken

unrestricted phrase-structure grammar

erlauben uneingeschränkte Regeln, d.h. keine Bedingungen für **x** und **y**

sind mächtiger als kontextsensitive Grammatiken

stehen an der Spitze der **Chomsky**-Hierarchie von PSGs

Transformationsgrammatiken

bestehen aus einer Regelbasis und Transformationsregeln, die Sätze auf wenige grammatikalische Grundformen reduzieren (transformieren), z.B.

Aktiv -> Passiv

Studenten lernen gerne schwierige Dinge.
-> Schwierige Dinge werden von Studenten gerne gelernt.

Augmented Transition Networks (ATNs)

"erweiterte Übergangsnetzwerke"

Parsing eines Satzes bedeutet, eine Folge von Ableitungen bzw. Regeln zu finden, die vom Startsymbol zum Satz führen.

Elementare Parsingalgorithmen

Top-down-Parser

arbeiten zielgerichtet, d.h. sie beginnen mit dem Startsymbol **S** und versuchen, durch eine Reihe von **Erweiterungen**, den Satz zu erzeugen

Bottom-up-Parser

arbeiten datengesteuert, d.h. sie beginnen mit dem Satz und suchen nach einer **Reduktion** auf das Startsymbol

Left-corner-Parser

verwenden einen **Mischstrategie**

Beispiel für eine einfache Grammatik

(G1): S -> NP VP

Ein Satz besteht aus Nominalphrase und Verbalphrase

(G2): NP -> n

Eine Nominalphrase besteht aus einem Nomen

(G3): VP -> v NP PP

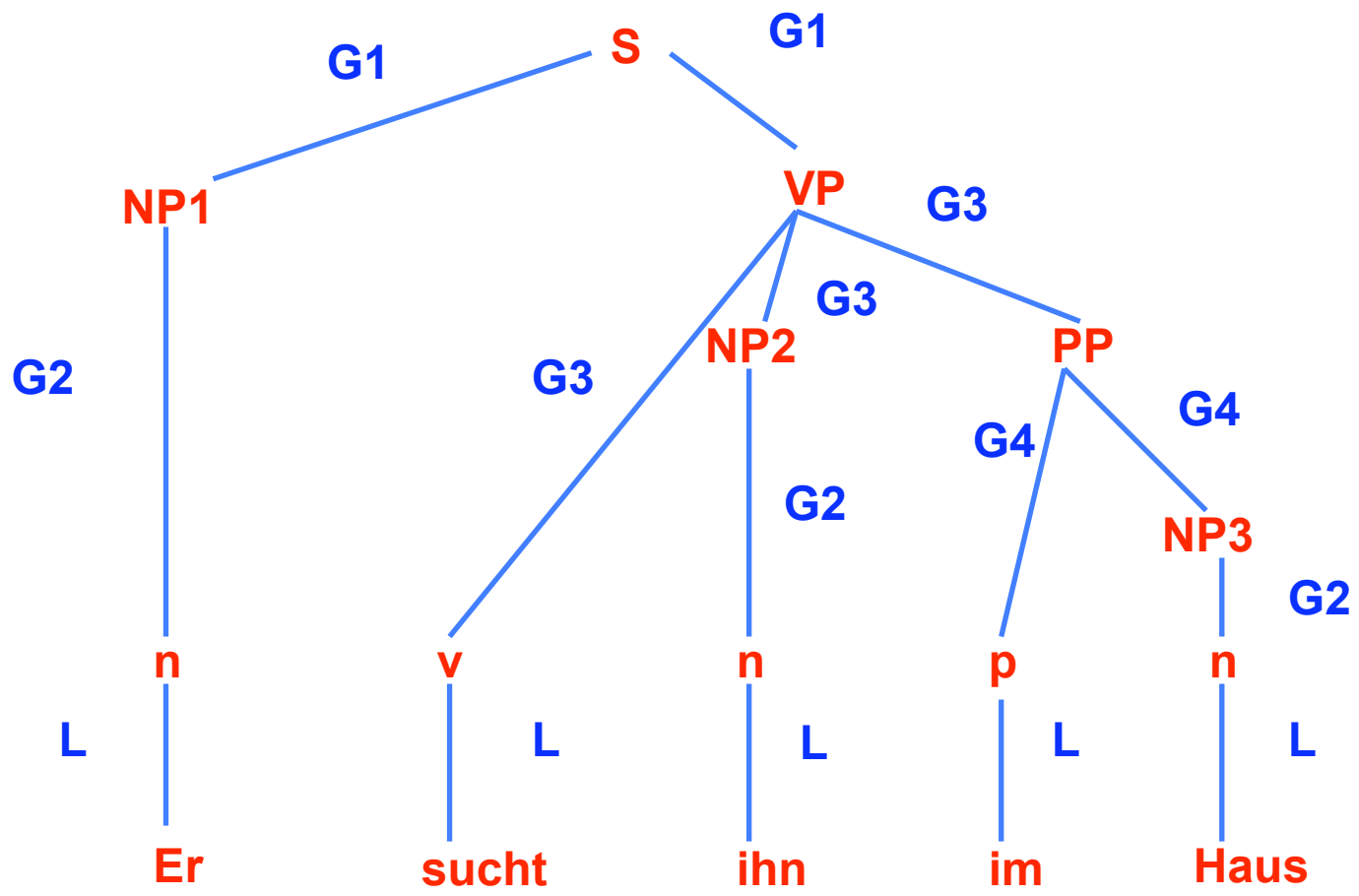
Eine Verbalphrase besteht aus einem Verb, einer Nominalphrase und einer Präpositionalphrase

(G4): PP -> p NP

Eine Präpositionalphrase besteht aus einer Präposition und einer Nominalphrase

Lexikon {er [n], haus [n], ihn [n], im [p], sucht [v]}

Satz: Er sucht ihn im Haus



Top-down-Parsing (vgl. Langer/Naumann, 410ff)

(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).

Umsetzung der Strategie Top-down

(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)

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

(6) **NP2 PP**
entferne das führende teminale Symbol v gemäß (P3)

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

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

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

(10) **NP3**
entferne das führende terminale Symbol gemäß (P3)

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

(12) beende das Parsing gemäß (P4)

Bottom-up-Parsing

(P1) Analysiere das nächste Wort.

(P2) Wenn es eine Regel gibt, deren rechte Seite mit den letzten Symbolen der Satzform übereinstimmt, dann ersetze sie durch die linke Seite der Regel und gehe zu (P2)

(P3) Gehe zu (P1)

Umsetzung der Strategie Bottom-up

(1) **n**
analysiere das nächste Wort gemäß (P1)

(2) **NP1**
ersetze n durch NP gemäß (P2) und (G2)

(3) **NP1 v**
(P2) trifft nicht zu, gehe gemäß (p3) zu (P1) und analysiere das nächste Wort v

(4) **NP1 v n**
(P2) trifft nicht zu, gehe gemäß (P3) zu (P1) und analysiere das nächste Wort n

(5) **NP1 v NP2**
ersetze n durch NP gemäß (P2) und (G2)

(6) **NP1 v NP2 p**
(P2) trifft nicht zu, gehe gemäß (P3) zu (P1) und analysiere das nächste Wort p

(7) **NP1 v NP2 p n**
(P2) trifft nicht zu, gehe gemäß (P3) zu (P1) und analysiere das nächste Wort n

(8) **NP1 v NP2 p NP3**
ersetze n durch NP gemäß (P2) und (G2)

(9) **NP1 v NP2 PP**
ersetze p NP durch PP gemäß (P2) und (G4)

(10) **NP1 VP**
ersetze v NP PP durch VP gemäß (P2) und (G3)

(11) **S**
ersetze NP VP durch S gemäß (P2) und (G1) und beende das Parsing, da S erreicht ist

Left-corner-Parsing

(P1) Die aktuelle Top-down-Erwartung ist S.

(P2) Analysiere das nächste Wort.

(P3) Suche eine Regel, deren linke Ecke (auf der rechten Seite) mit der gefundenen Kategorie übereinstimmt. Die Kategorien der rechten Regelseite (außer der linken Ecke) werden verwendet, um die Top-down-Erwartung zu aktualisieren.

(P4) Gehe zu (P2).

Umsetzung der Strategie Left-corner

(0) **S**
die aktuelle Top-down-Erwartung ist S gemäß (P1)

(1) **n**
analysiere das nächste Wort gemäß (P2)

(2) **NP1**
ersetze n durch NP gemäß (G2)

(3) **S / VP**
NP entspricht der linken Ecke von (G1), die aktuelle Top-down-Erwartung ist gemäß (P3) VP

(4) **v**
analysiere das nächste Wort gemäß (P2)

(5) **VP / NP2 PP**
v entspricht der linken Ecke von (G3), die aktuelle Top-down-Erwartung ist gemäß (P3) NP PP

(6) **n**
analysiere das nächste Wort gemäß (P2)

(7) **NP2**
ersetze n durch NP gemäß (G3)

(8) **p**
analysiere das nächste Wort gemäß (P2)

(9) **PP / NP3**
p entspricht der linken Ecke von (G4), die aktuelle Top-down-Erwartung ist gemäß (P3) NP

(10) **n**
analysiere das nächste Wort gemäß (P2)

(11) **NP3**
ersetze n durch NP gemäß (G2) und beende das Parsing

Grenzen der elementaren Parsing-Algorithmen

arbeiten nur für einfache Syntax, d.h. für jeden Schritt darf nur eine Regel existieren

erlauben keine Analysealternativen, z.B. Umkehr bei Misserfolg und neuer Versuch

speichern keine Teilergebnisse

sog. Chart-Parser und LR-Parser lösen eine oder mehrere dieser Einschränkungen

Übung 5

Gegeben sei folgende Grammatik

<S> ::= <NP> <VP>

<NP> ::= <*N> | <*PRO>

<VP> ::= <*V> <NP>

und folgender Satz

Ich esse Käse.

Geben sie die Parsingverläufe für Top-down und Bottom-up an.

Wie sehen die Parsingverläufe für folgenden Satz aus?

Ich esse grünen Käse.

Allen, James: Natural Language Understanding. Redwood 1995.

Duden: Grammatik der deutschen Gegenwartssprache. 6. Aufl. Mannheim 1998.

Grishman, Ralph: Computational Linguistics: an Introduction. Cambridge 1986.

Hausser, Roland: Grundlagen der Computerlinguistik: Mensch-Maschine-Kommunikation in natürlicher Sprache. Berlin u.a. 2000.

Nohr, Holger: Automatische Indexierung: Einführung in betriebliche Verfahren, Systeme und Anwendungen. Potsdam 2001. (Materialien zur Information und Dokumentation; Band 13)

Sprachverarbeitung. In: Görz, Günther (Hrsg.): Einführung in die künstliche Intelligenz, 2. Auflage, Bonn u.a.1995, S. 361-557.

Volmert, Johannes (Hrsg.): Grundkurs Sprachwissenschaft: eine Einführung in die Sprachwissenschaft für Lehramtsstudiengänge. München 1995.