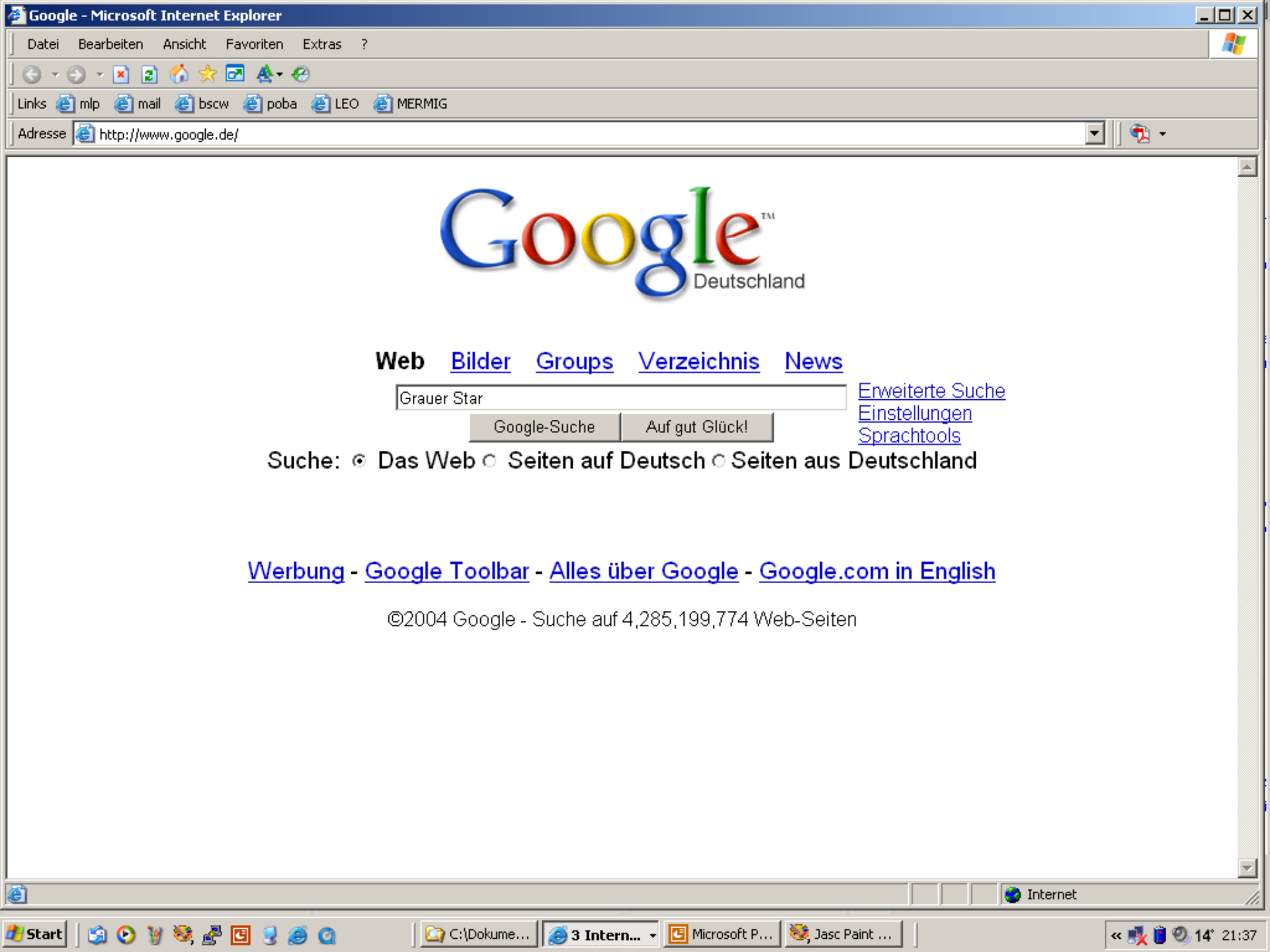


# Verbesserung der Recherche in medizinischen Textkollektionen durch Wortstamm-basierte Indexierung

**Stefan Schulz**

Abteilung Medizinische Informatik,  
Universitätsklinikum Freiburg





**Web** [Bilder](#) [Groups](#) [Verzeichnis](#) [News](#)

Grauer Star

Google-Suche Auf gut Glück!

[Erweiterte Suche](#)  
[Einstellungen](#)  
[Sprachtools](#)

Suche:  Das Web  Seiten auf Deutsch  Seiten aus Deutschland

[Werbung](#) - [Google Toolbar](#) - [Alles über Google](#) - [Google.com in English](#)

©2004 Google - Suche auf 4,285,199,774 Web-Seiten

Anfrage an ein Textretrieval-System (Suchmaschine):  
**„Grauer Star“**

Anfrage an ein Textretrieval-System (Suchmaschine):

„**Grauer Star**“

Suchmaschine findet u.a. nicht relevante Dokumente:

### **Vögel und Merkmale**

Unsere Vögel und ihre Merkmale (die Namen in Klammern sind von den Wölflingen, die die Merkmale zusammengetragen haben). Blaumeise. ...

**Star**. ... **grauer** Kopf. (Michi). ...

[www.pfadfinder-traustadt.de/wir/meute/projekte/voegel/voegelundmerkmale.htm](http://www.pfadfinder-traustadt.de/wir/meute/projekte/voegel/voegelundmerkmale.htm) - 13k

### **Vogelgeschichten -**

Der kleine **Star**... Keine Katze, kein Hund, kein älterer **grauer** Mann ... auf dem Friedhof auskannte, bat die anderen Vögel und auch ... Eines Tages traf der Star zwei kleine Eichhörnchen ...

[www.tiergeschichten.de/voegel/derkleinestar.htm](http://www.tiergeschichten.de/voegel/derkleinestar.htm) - 21k -

Anfrage an ein Textretrieval-System (Suchmaschine):

**„Grauer Star“**

Suchmaschine findet u.a. nicht relevante Dokumente:

### **Vögel und Merkmale**

Unsere Vögel und ihre Merkmale (die Namen in Klammern sind von den Wölflingen, die die Merkmale zusammengetragen haben). Blaumeise. ...

**Star.** ... **grauer** Kopf. (Michi). ...

[www.pfadfinder-traustadt.de/wir/meute/projekte/voegel/voegelundmerkmale.htm](http://www.pfadfinder-traustadt.de/wir/meute/projekte/voegel/voegelundmerkmale.htm) - 13k

### **Vogelgeschichten -**

Der kleine **Star**... Keine Katze, kein Hund, kein älterer **grauer** Mann ... auf dem Friedhof auskannte, bat die anderen Vögel und auch ... Eines Tages traf der Star zwei kleine Eichhörnchen ...

[www.tiergeschichten.de/voegel/derkleinestar.htm](http://www.tiergeschichten.de/voegel/derkleinestar.htm) - 21k -

Suchmaschine findet relevante Dokumente nicht:

### **Patienteninformationen/Vorderer Abschnitt des Auges/Der graue Star ...**

... Patienteninformationen/Vorderer Abschnitt des Auges/Der graue Star (Katarakt), Druckversion. ... Der Graue Star (Katarakt). ... Wie wird der Graue Star behandelt? ...

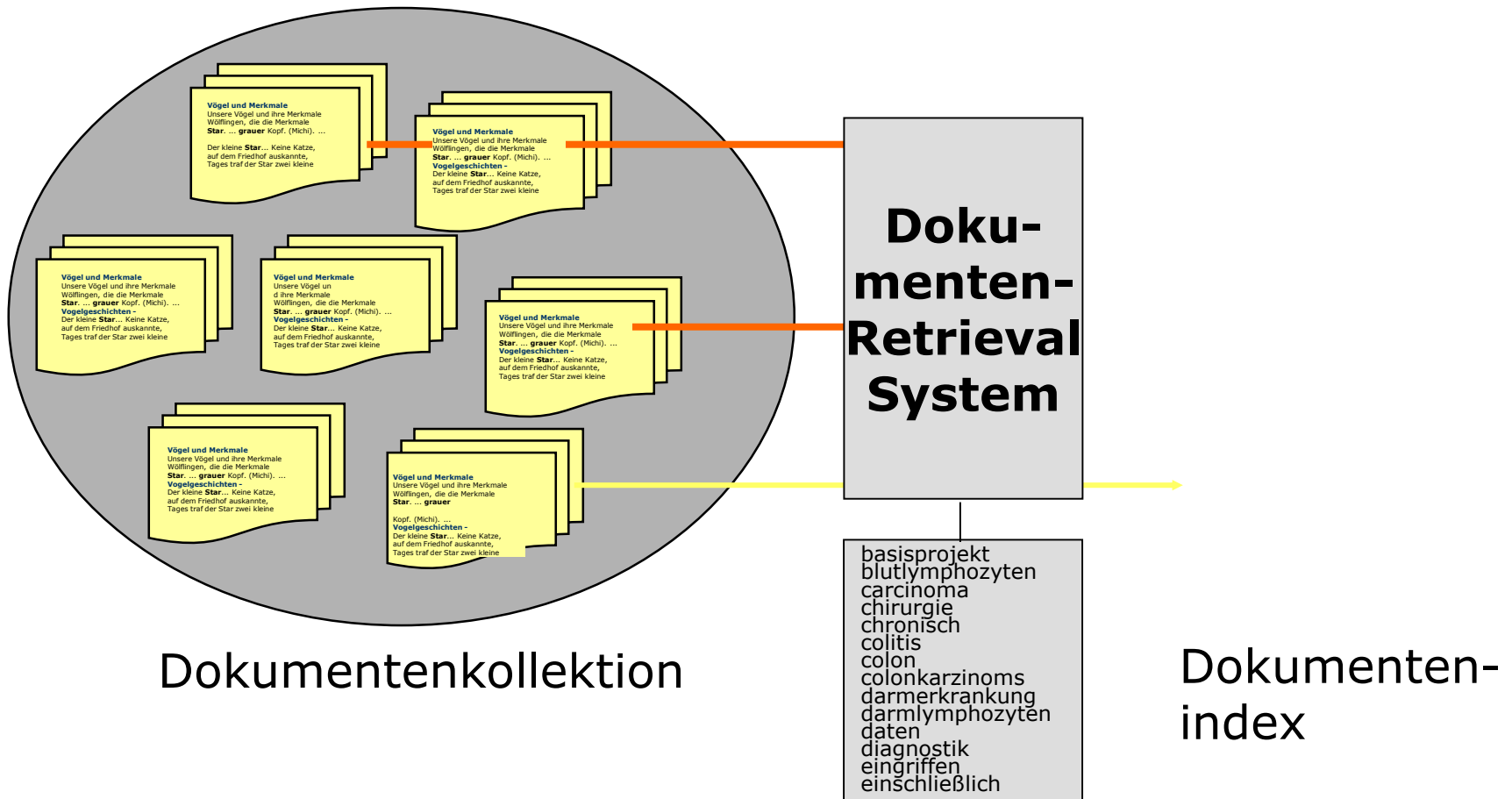
[www.uniklinikum-giessen.de/augen/katarakt.html](http://www.uniklinikum-giessen.de/augen/katarakt.html) - 26k

### **Erhöhtes Katarakt-Risiko auch bei inhalierten Steroiden**

Bad Drug News -- Erhöhtes Katarakt-Risiko auch bei inhalierten Steroiden. ... (UPM) Eine Therapie mit Steroiden bedeutet ein erhöhtes Katarakt-Risiko. ...

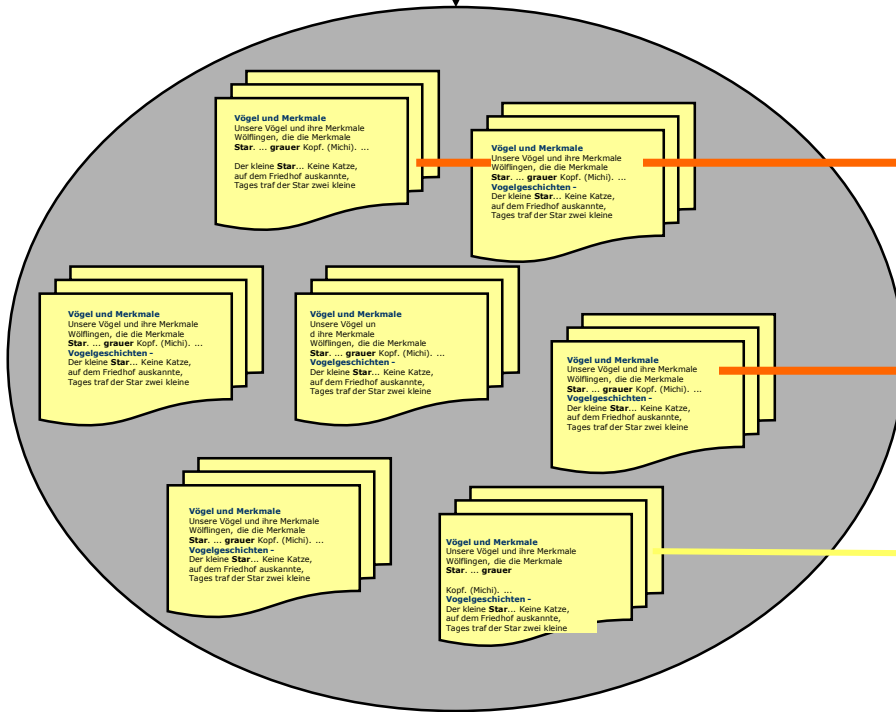
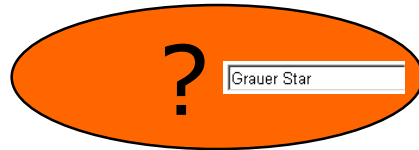
[www.infomed.org/bad-drug-news/bdn115.html](http://www.infomed.org/bad-drug-news/bdn115.html) -

# Textretrieval

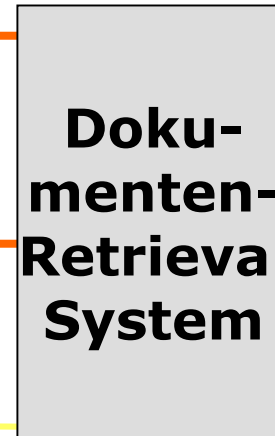


# Textretrieval

Anfrage  
("query")



Dokumentensammlung



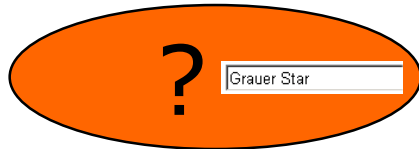
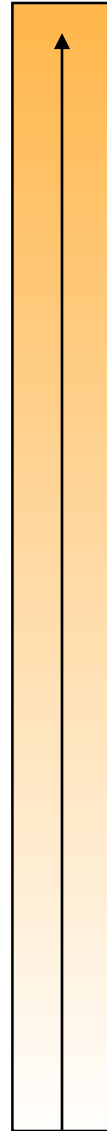
basisprojekt  
blutlymphozyten  
carcinoma  
chirurgie  
chronisch  
colitis  
colon  
colonkarzinoms  
darmerkrankung  
darmlymphozyten  
daten  
diagnostik  
eingriffen  
einschließlich

Dokumenten-  
index

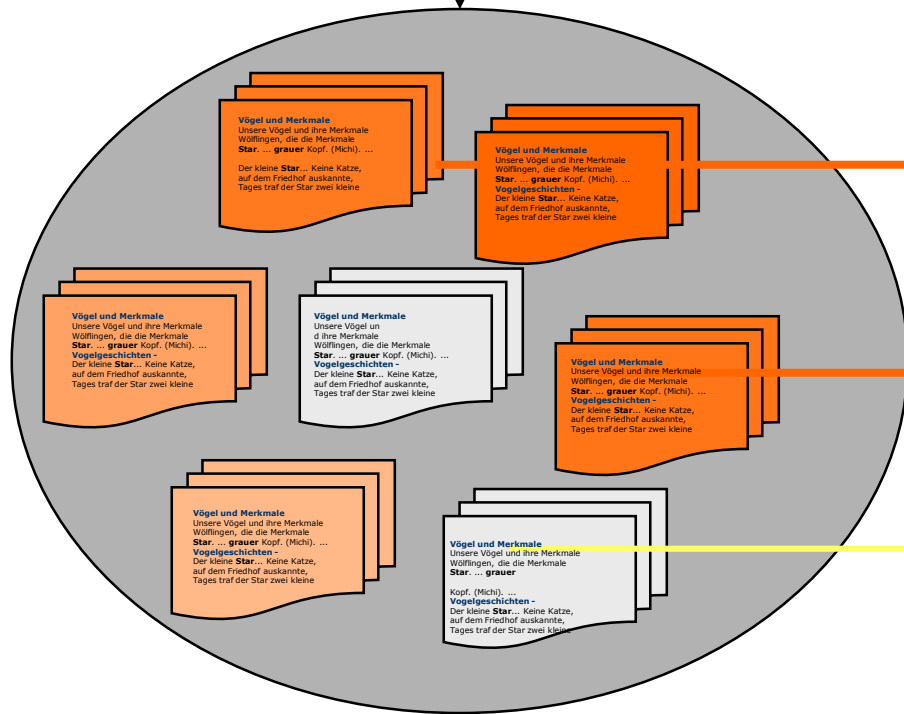


# Textretrieval

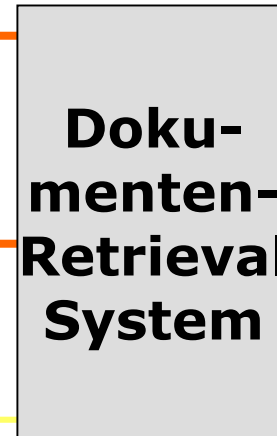
Relevanz



Anfrage ("query")



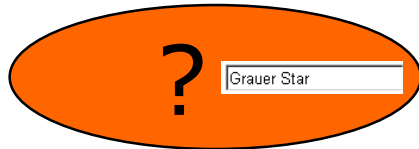
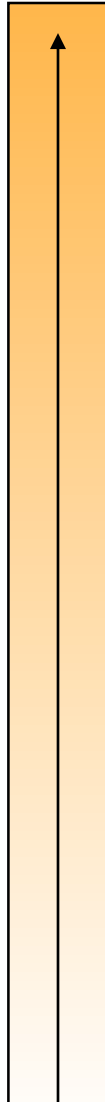
Dokumentenkollektion



Dokumenten-index

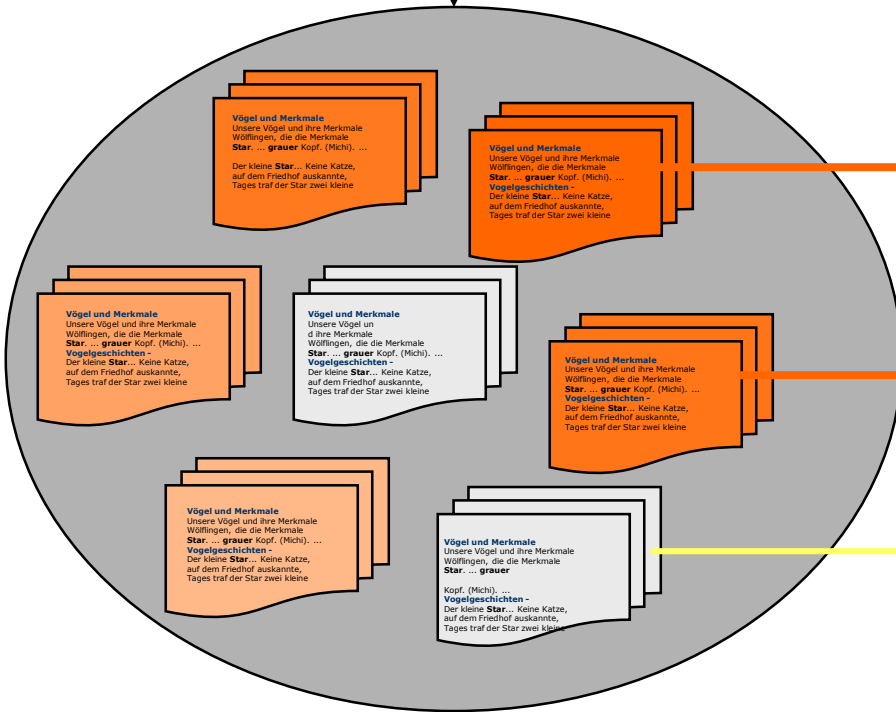
# Textretrieval

Relevanz

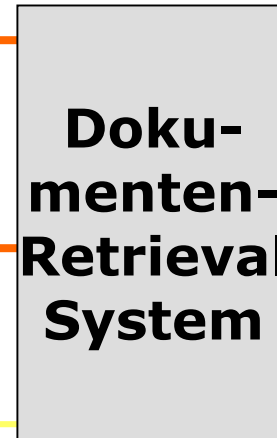


Anfrage ("query")

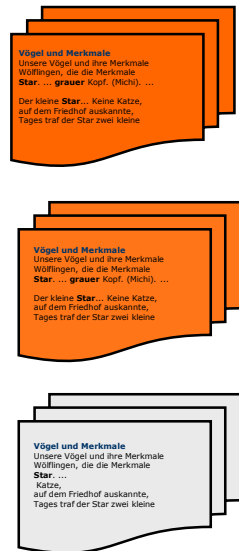
Ergebnisse der Recherche



Dokumentenkollektion



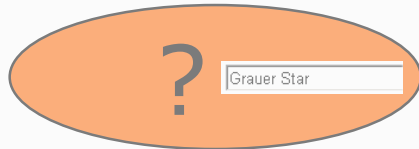
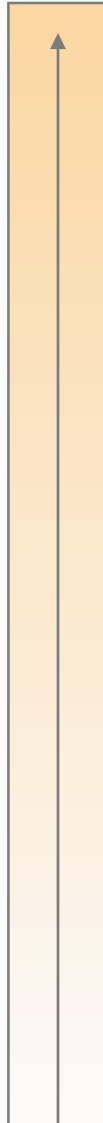
- basisprojekt
- blutlymphozyten
- carcinoma
- chirurgie
- chronisch
- colitis
- colon
- colonkarzinoms
- darmerkrankung
- darmlymphozyten
- daten
- diagnostik
- eingriffen
- einschließlich



Dokumentenindex

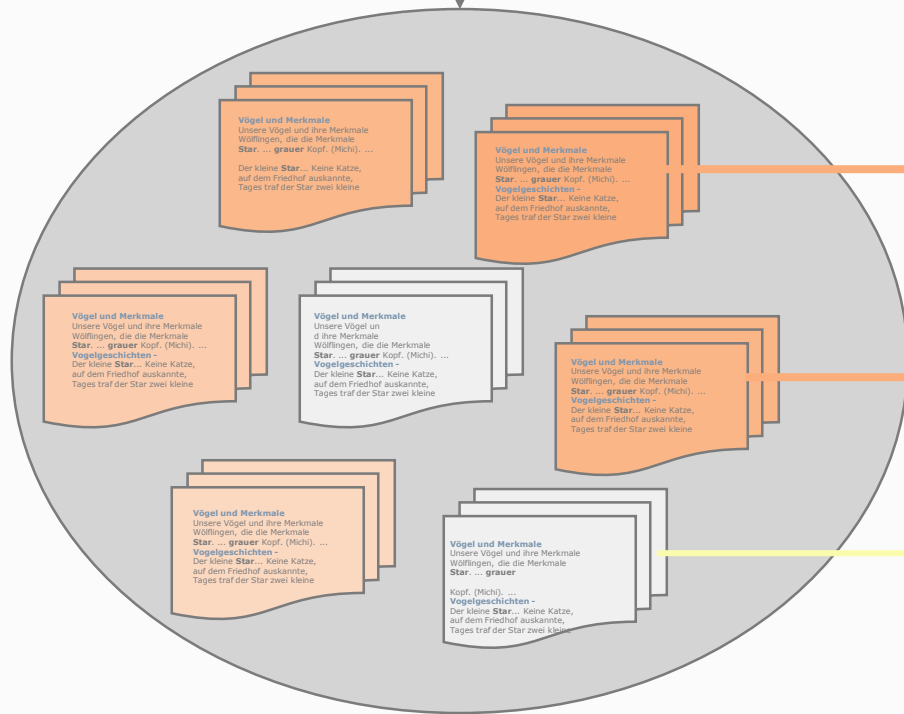
# Textretrieval

Relevanz



Anfrage ("query")

Ergebnisse der Recherche



Dokumentenkollektion



- basisprojekt
- blutlymphozyten
- carcinoma
- chirurgie
- chronisch
- colitis
- colon
- colonkarzinoms
- darmerkrankung
- darmlymphozyten
- daten
- diagnostik
- eingriffen
- einschließlich



Dokumentenindex

# Automatische Indexierung: Wortindex

AG Gastroenterologie - Microsoft Internet Explorer

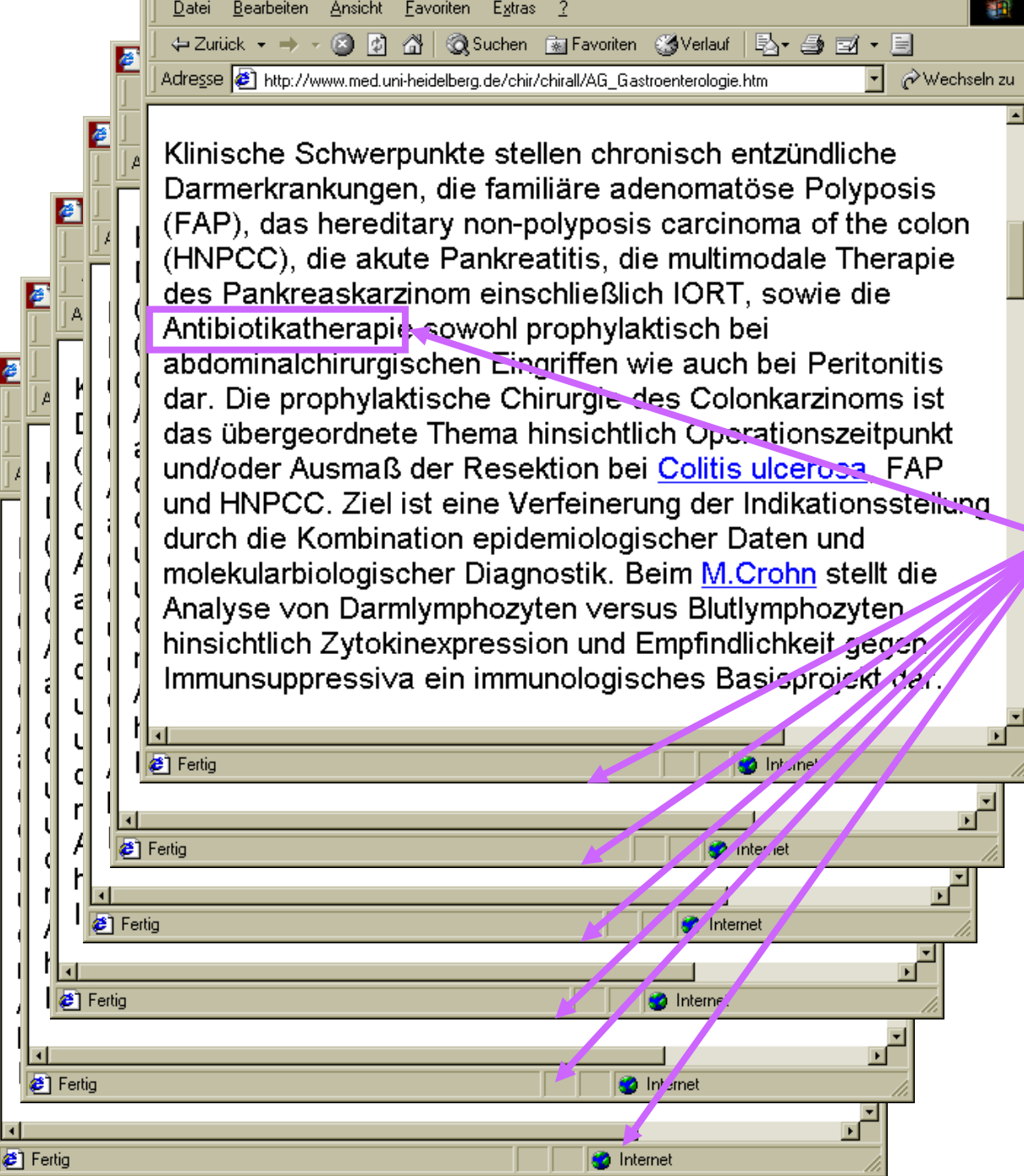
Adresse [http://www.med.uni-heidelberg.de/chir/chirall/AG\\_Gastroenterologie.htm](http://www.med.uni-heidelberg.de/chir/chirall/AG_Gastroenterologie.htm)

Klinische Schwerpunkte stellen chronisch entzündliche Darmerkrankungen, die familiäre adenomatöse Polyposis (FAP), das hereditary non-polyposis carcinoma of the colon (HNPCC), die akute Pankreatitis, die multimodale Therapie des Pankreaskarzinom einschließlich IORT, sowie die Antibiotikatherapie sowohl prophylaktisch bei abdominalchirurgischen Eingriffen wie auch bei Peritonitis dar. Die prophylaktische Chirurgie des Colonkarzinoms ist das übergeordnete Thema hinsichtlich Operationszeitpunkt und/oder Ausmaß der Resektion bei Colitis ulcerosa, FAP und HNPCC. Ziel ist eine Verfeinerung der Indikationsstellung durch die Kombination epidemiologischer Daten und molekularbiologischer Diagnostik. Beim M.Crohn stellt die Analyse von Darmlymphozyten versus Blutlymphozyten hinsichtlich Zytokinexpression und Empfindlichkeit gegen Immunsuppressiva ein immunologisches Basisprojekt dar.

abdominalchirurgischen  
adenomatöse  
akute  
analyse  
antibiotikatherapie  
ausmaß  
basisprojekt  
blutlymphozyten  
carcinoma  
chirurgie  
chronisch  
colitis  
colon  
colonkarzinoms  
darmerkrankungen  
darmlymphozyten  
daten  
diagnostik  
eingriffen  
einschließlich  
empfindlichkeit  
entzündliche  
epidemiologischer

Klinische Schwerpunkte stellen chronisch entzündliche Darmerkrankungen, die familiäre adenomatöse Polyposis (FAP), das hereditary non-polyposis carcinoma of the colon (HNPCC), die akute Pankreatitis, die multimodale Therapie des Pankreaskarzinom einschließlich IORT, sowie die Antibiotikatherapie sowohl prophylaktisch bei abdominalchirurgischen Eingriffen wie auch bei Peritonitis dar. Die prophylaktische Chirurgie des Colonkarzinoms ist das übergeordnete Thema hinsichtlich Operationszeitpunkt und/oder Ausmaß der Resektion bei Colitis ulcerosa, FAP und HNPCC. Ziel ist eine Verfeinerung der Indikationsstellung durch die Kombination epidemiologischer Daten und molekularbiologischer Diagnostik. Beim M.Crohn stellt die Analyse von Darmlymphozyten versus Blutlymphozyten hinsichtlich Zytokinexpression und Empfindlichkeit gegen Immunsuppressiva ein immunologisches Basisprojekt dar.

- abdominalchirurgischen
- adenomatöse
- akute
- analyse
- antibiotikatherapie
- ausmaß
- basisprojekt
- blutlymphozyten
- carcinoma
- chirurgie
- chronisch
- colitis
- colon
- colonkarzinoms
- darmerkrankungen
- darmlymphozyten
- daten
- diagnostik
- eingriffen
- einschließlich
- empfindlichkeit
- entzündliche
- epidemiologischer



# Indexierung auf Wort-Ebene

## Probleme:

- Linguistische Phänomene erschweren medizinisches Text-Retrieval, z.B.
- Morphologische Prozesse:
  - *Flexion*: *Leukozyt* <> *Leukozyten*, *Ulcus* <> *ulcera*
  - *Derivation*: *Leukozyt* <> *leukozytär*
  - *Komposition*: *Leuk|ämie*, *Rechts|herz|insuffizienz*
- Orthographische Variation
  - *Karzinom* <> *Carcinom* <> *Carzinom*
- Synonymie, Variationen der Rechtschreibung:
  - *Ascorbinsäure* <> *Vitamin C*, *Haut* <> *Cutis*

# Lösungsansatz:

## Subwort-Index statt Wort-Index

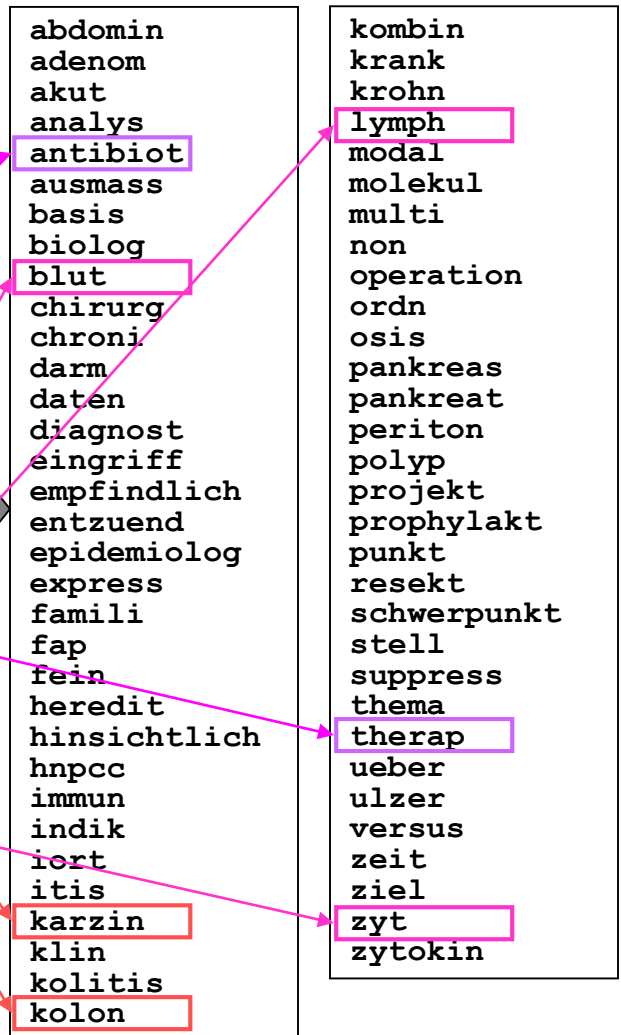
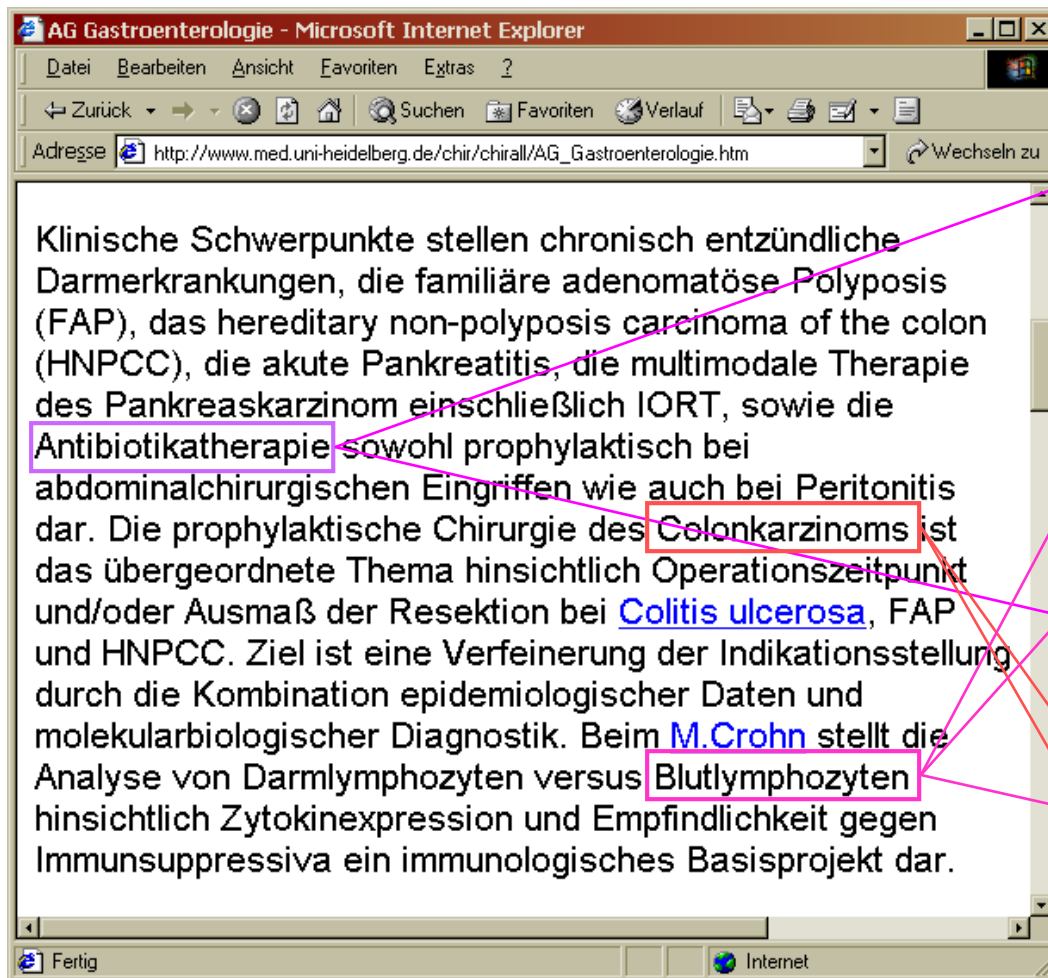
- Subwörter sind atomare Begriffs- oder linguistische Einheiten:
  - Stämme: verletz, entzünd, magen, schleimhaut
  - Präfixe: ab-, an-, anti-, ge-, hervor-, hyper-
  - Suffixe: -abel, -bar, -haft, -ion, -itis
  - Infixe: -o-, -s-
- Synonyme Subwörter werden in Synonymklassen gruppiert:
  - **kqxqqk** = {nephr, niere, kidney}
  - **kqxqqk** = {leber, hepat, liver}

# Ressourcen

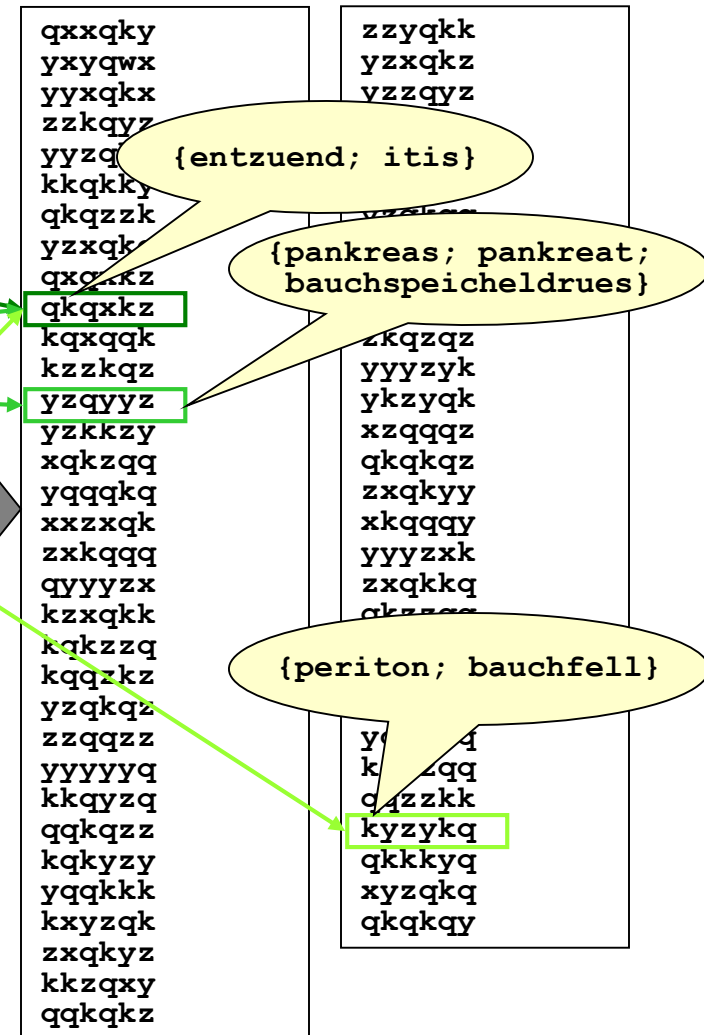
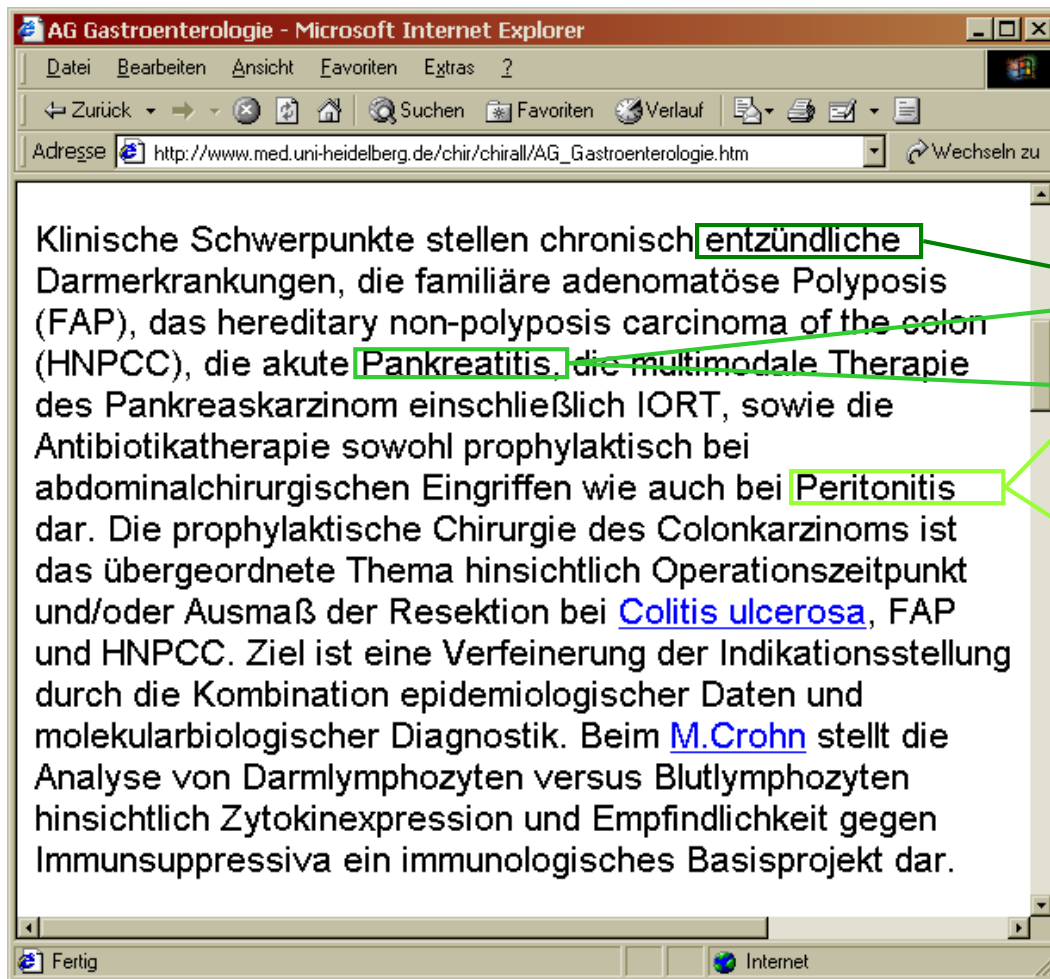
- Subwort-Lexikon:
  - Organisiert und klassifiziert medizinspezifische Subwörter und Affixe in mehreren Sprachen (derzeit Deutsch, Englisch, Portugiesisch, ca. 25.000 Einträge), Spanisch, Französisch, Schwedisch im Aufbau
- Subwort-Thesaurus:
  - Gruppiert synonyme Lexikoneinträge
- Morphosyntaktischer Parser:
  - Extrahiert aus Texten Subwörter und ordnet ihnen Synonymklassen – IDs zu



# Indexierung durch Subwörter



# Indexierung durch Subwort – Synonymklassen-IDs



# Evaluation

Wissenschaftliche Fragestellung:

Verbessert ein automatisch erstellter  
Subwort-Index die Recherche in  
medizinischen Dokumentenbeständen ?

# Textretrievalsysteme: Evaluationsmethodik

## ■ Kenngrößen:

$$precision = \frac{n_{\text{gefundene+relevante Dokumente}}}{n_{\text{gefundene Dokumente}}}$$

$$recall = \frac{n_{\text{gefundene+relevante Dokumente}}}{n_{\text{relevante Dokumente}}}$$

## ■ Precision/Recall-Diagramme bei geranktem Output

Beispiel: 25 Dokumente, 8 relevant

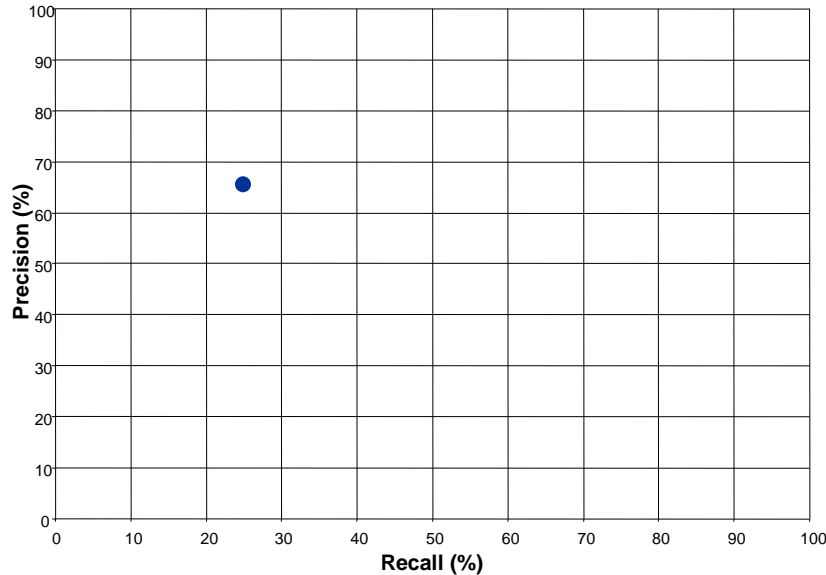
Anfrage X



<i>precision</i> = 67%	Dokument 05
<i>recall</i> = 25%	Dokument 16
	Dokument 21

Dokument 22  
Dokument 02  
Dokument 25  
Dokument 20  
Dokument 10  
Dokument 07  
Dokument 18  
Dokument 04  
Dokument 12  
Dokument 11  
Dokument 24  
Dokument 15  
Dokument 09  
Dokument 17  
Dokument 08  
Dokument 19  
Dokument 13  
Dokument 03  
Dokument 14  
Dokument 23  
Dokument 01  
Dokument 06

# Textretrievalsysteme: Evaluationsmethodik



## ■ Precision/Recall-Diagramme bei geranktem Output

Beispiel: 25 Dokumente, 8 relevant

Anfrage X



*precision* =

*recall* =

Dokument 05

Dokument 16

Dokument 21

Dokument 22

Dokument 02

Dokument 25

Dokument 20

Dokument 10

Dokument 07

Dokument 18

Dokument 04

Dokument 12

Dokument 11

Dokument 24

Dokument 15

Dokument 09

Dokument 17

Dokument 08

Dokument 19

Dokument 13

Dokument 03

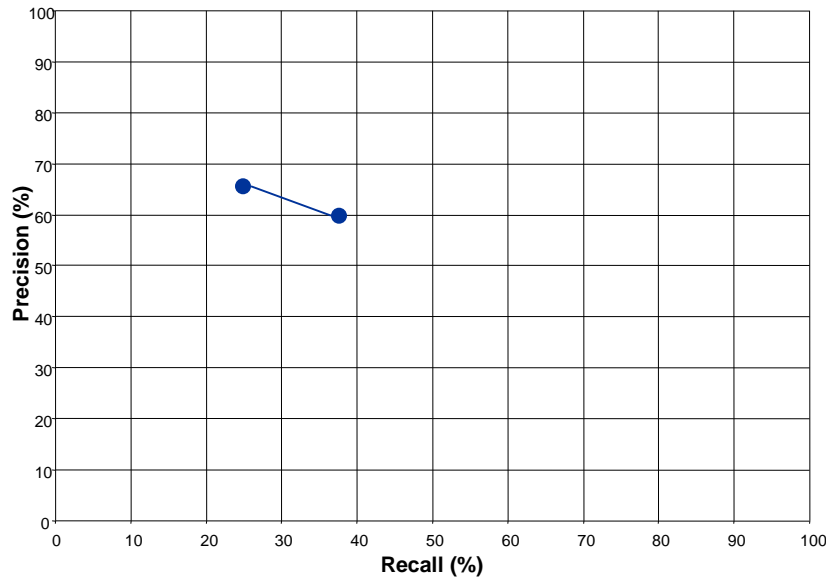
Dokument 14

Dokument 23

Dokument 01

Dokument 06

# Textretrievalsysteme: Evaluationsmethodik



## ■ Precision/Recall-Diagramme bei geranktem Output

Beispiel: 25 Dokumente, 8 relevant

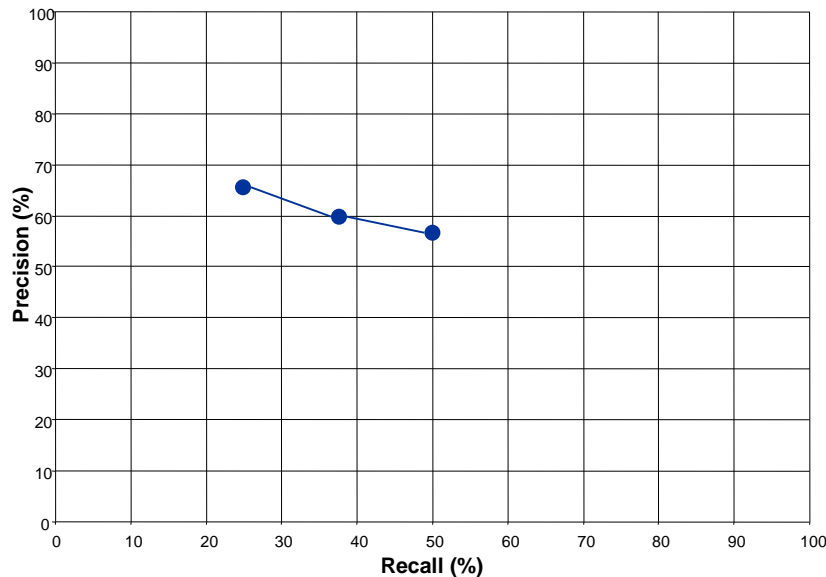
Anfrage X

*precision* = 60%

*recall* = 38%

↓  
Dokument 05  
Dokument 16  
Dokument 21  
Dokument 22  
Dokument 02  
Dokument 25  
Dokument 20  
Dokument 10  
Dokument 07  
Dokument 18  
Dokument 04  
Dokument 12  
Dokument 11  
Dokument 24  
Dokument 15  
Dokument 09  
Dokument 17  
Dokument 08  
Dokument 19  
Dokument 13  
Dokument 03  
Dokument 14  
Dokument 23  
Dokument 01  
Dokument 06

# Textretrievalsysteme: Evaluationsmethodik



## ■ Precision/Recall-Diagramme bei geranktem Output

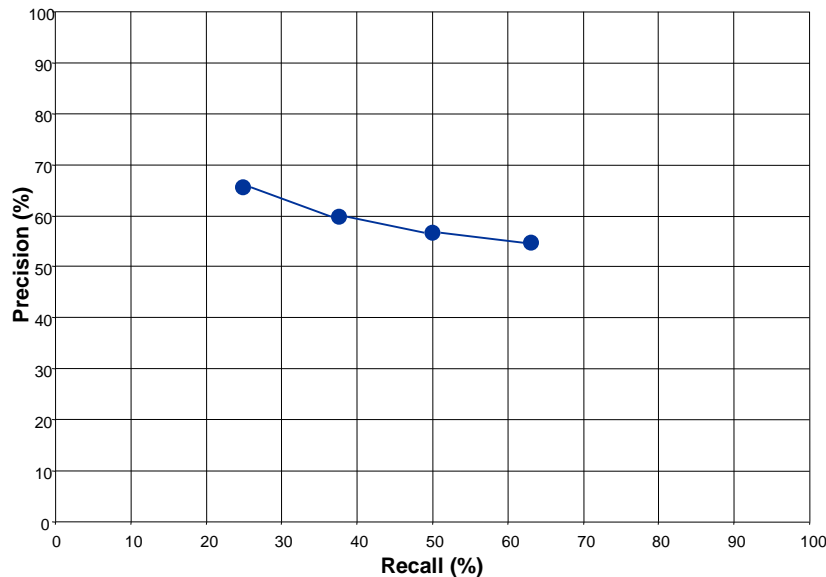
Beispiel: 25 Dokumente, 8 relevant

Anfrage X

*precision* = 57%  
*recall* = 50%

↓  
Dokument 05  
Dokument 16  
Dokument 21  
Dokument 22  
Dokument 02  
Dokument 25  
Dokument 20  
Dokument 10  
Dokument 07  
Dokument 18  
Dokument 04  
Dokument 12  
Dokument 11  
Dokument 24  
Dokument 15  
Dokument 09  
Dokument 17  
Dokument 08  
Dokument 19  
Dokument 13  
Dokument 03  
Dokument 14  
Dokument 23  
Dokument 01  
Dokument 06

# Textretrievalsysteme: Evaluationsmethodik



## ■ Precision/Recall-Diagramme bei geranktem Output

Beispiel: 25 Dokumente, 8 relevant

Anfrage X

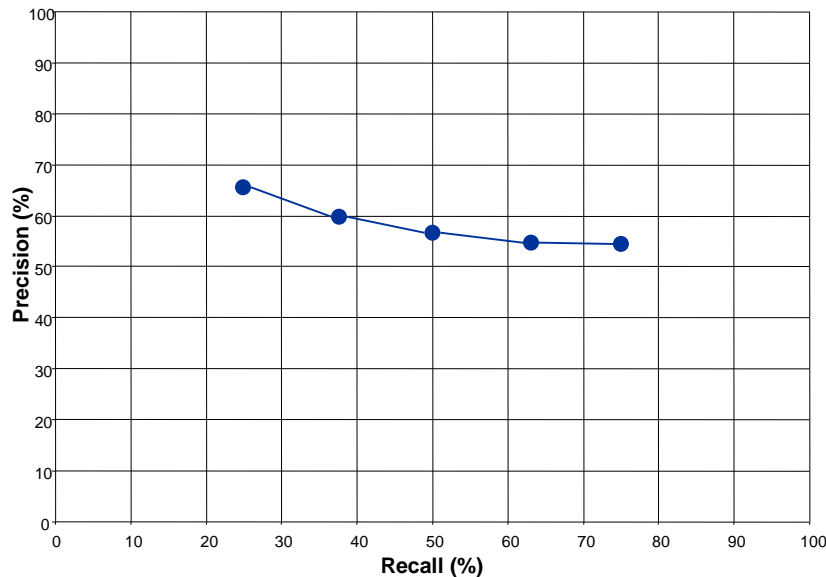
*precision* = 55%  
*recall* = 63%

↓

Dokument 05  
Dokument 16  
Dokument 21  
Dokument 22  
Dokument 02  
Dokument 25  
Dokument 20  
Dokument 10  
Dokument 07  
Dokument 18  
Dokument 04  
Dokument 12  
Dokument 11  
Dokument 24  
Dokument 15  
Dokument 09  
Dokument 17  
Dokument 08  
Dokument 19  
Dokument 13  
Dokument 03  
Dokument 14  
Dokument 23  
Dokument 01  
Dokument 06



# Textretrievalsysteme: Evaluationsmethodik



## ■ Precision/Recall-Diagramme bei geranktem Output

Beispiel: 25 Dokumente, 8 relevant

Anfrage X



*precision* = 54%  
*recall* = 75%

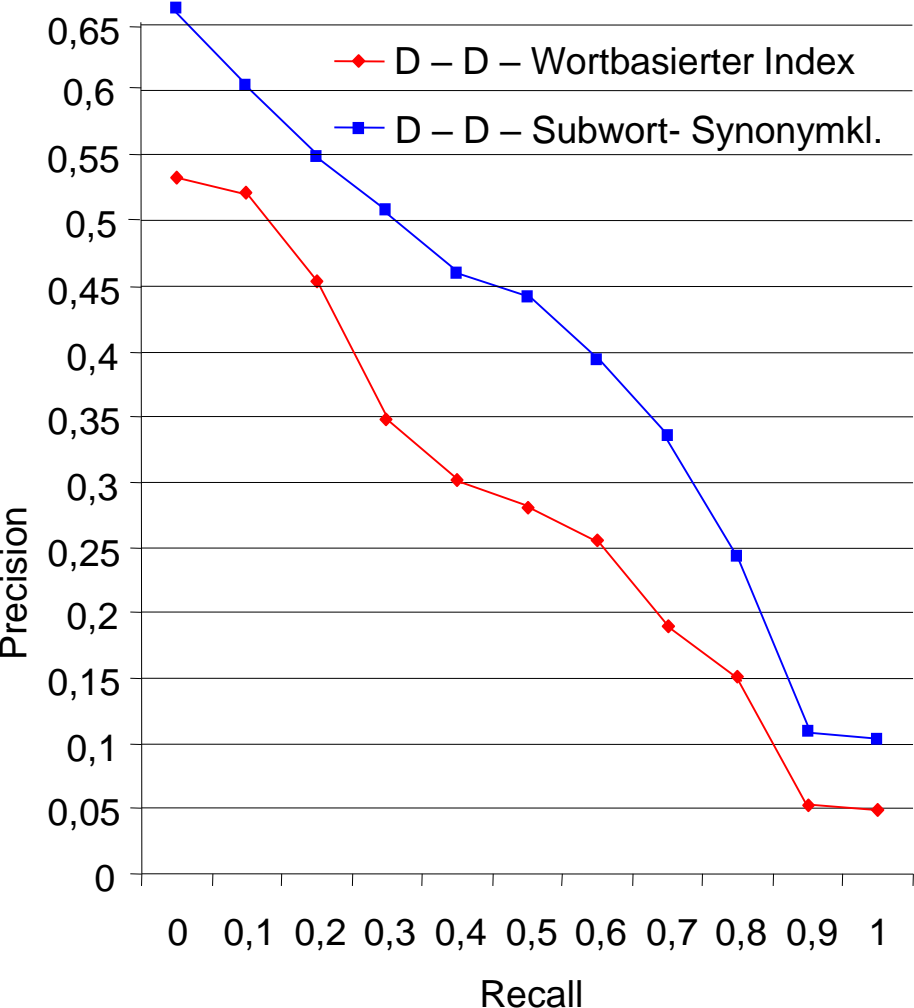
Dokument 05  
Dokument 16  
Dokument 21  
Dokument 22  
Dokument 02  
Dokument 25  
Dokument 20  
Dokument 10  
Dokument 07  
Dokument 18  
Dokument 04  
Dokument 12  
Dokument 11  
Dokument 24  
Dokument 15  
Dokument 09  
Dokument 17  
Dokument 08  
Dokument 19  
Dokument 13  
Dokument 03  
Dokument 14  
Dokument 23  
Dokument 01  
Dokument 06

# Evaluationsszenarien

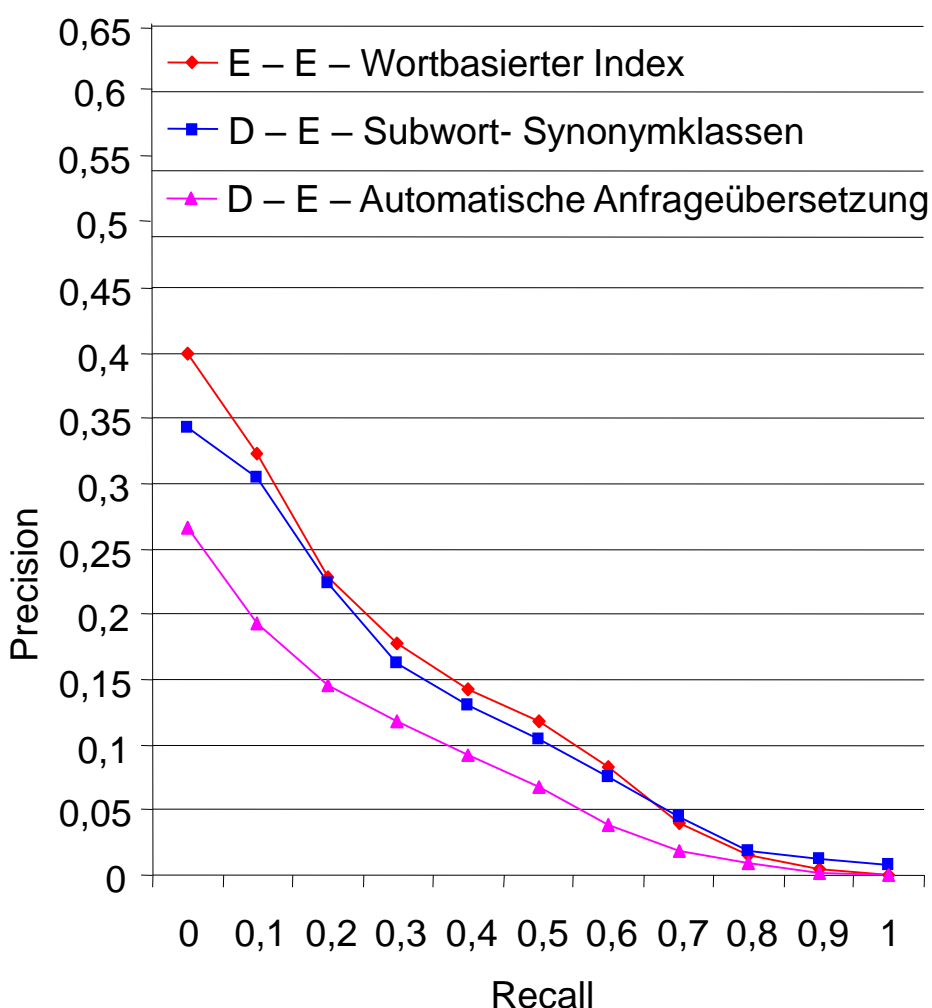
	<b>Szenario 1</b>	<b>Szenario 2</b>
Sprachen	D: Deutsch Q: Deutsch	D: Englisch Q: Deutsch, Englisch
Dokumente	MSD-Manual ( $ D  = 5.500$ )	MEDLINE-Abstracts ( $ D  = 233.000$ )
Anfragen	$ Q  = 25$ (nach IMPP-Fragen durch Medizinstudenten, Uni FR)	$ Q  = 106$ (Oregon Health Science Univ.) Übersetzung durch Medizin- studenten ins Deutsche
Goldstandard: $D \otimes Q \rightarrow \{rel, n.rel\}$	Relevanzurteile durch Einzelbewertung Medizinstudenten, Uni FR	Relevanzurteile durch MeSH-vermittelte Medline-Anfragen und manuelle Nachbearbeitung durch med. Dokumentare

# Ergebnisse

## Szenario 1



## Szenario 2



# Folgerung

- Indexierung mit Subwort-Synonymklassen verbessert das Retrieval in medizinischen Textkollektionen
- Nachweis für sprachinternes Retrieval (deutsch-deutsch) und für sprachübergreifendes Retrieval (deutsch-englisch)
- Abdeckungsgrad und Qualität des Lexikons von entscheidender Bedeutung

# Stand des Projekts

## ■ Finanzierung:

- DFG – Projekt KoMoDoRe
- BMBF – Internationales Büro: Wissenschaftleraustausch
- EU – SemanticMining Network of Excellence

## ■ Partner:

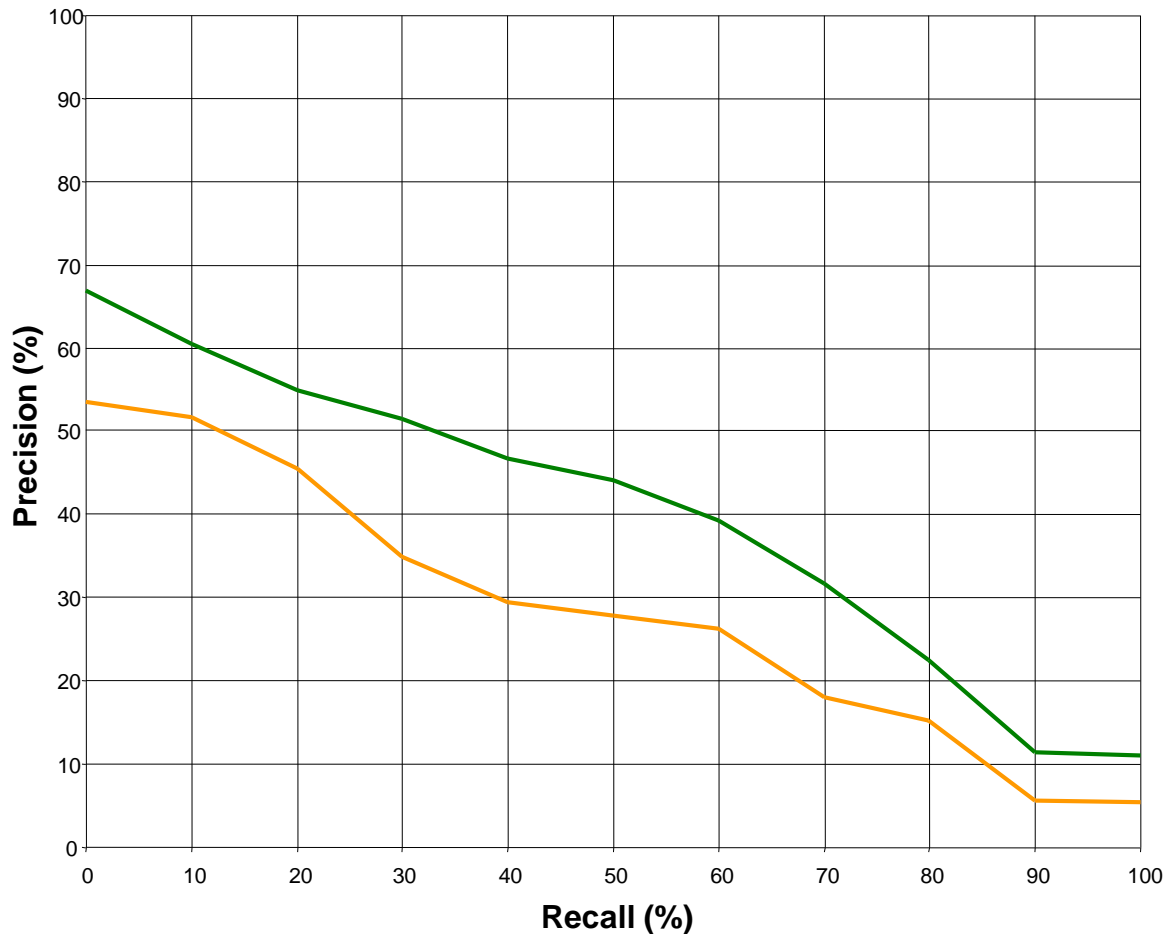
- Universitätsklinikum Freiburg, Medizinische Informatik (Projektleitung)
- Universität Jena, Abteilung Computerlinguistik
- Katholische Universität Paraná, Curitiba, Brasilien
- Sahlgrenska Universitätsklinikum Göteborg, Schweden
- Universität Göteborg, Schwedische Sprachwissenschaft
- Kantonshospital Genf, Medizinische Informatik (Schweiz)

[www.morphosaurus.de](http://www.morphosaurus.de)





# Ergebnisse: Szenario 1



Precision-Recall-Diagramm:  
- Precisionwerte an fixen  
Recall-Leveln durch Interpolation  
- Mittelwert aus 25 Messreihen

Wortindex

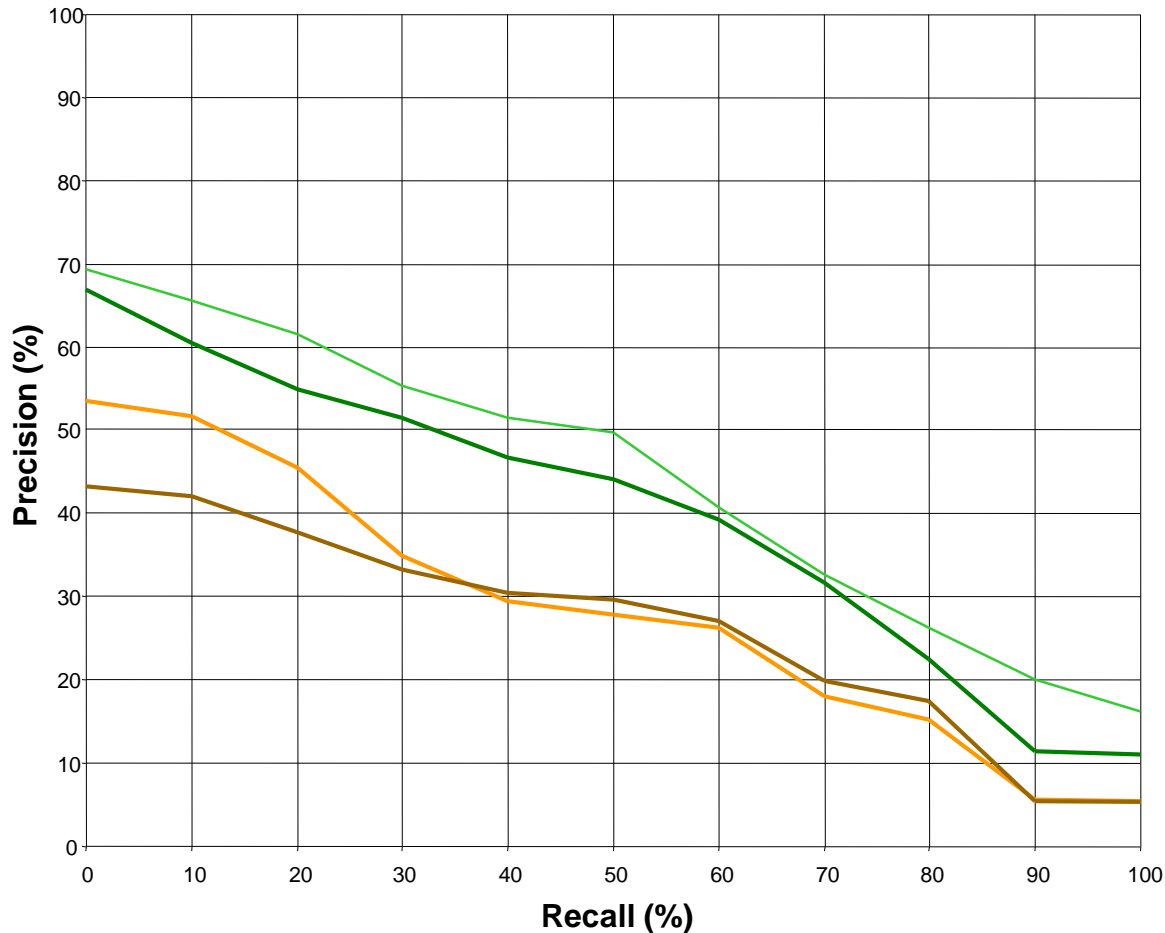
Subwort- Synonymklassen



# Evaluation: Retrievalszenarien

- Suchmaschine: AltaVista™, lokal installiert
- Szenarien
  - 1. Wortindex
  - 2. Wortindex mit Stammformenreduktion
  - 3. Subwordindex ohne Semantik
  - 4. Subwordindex mit Semantik

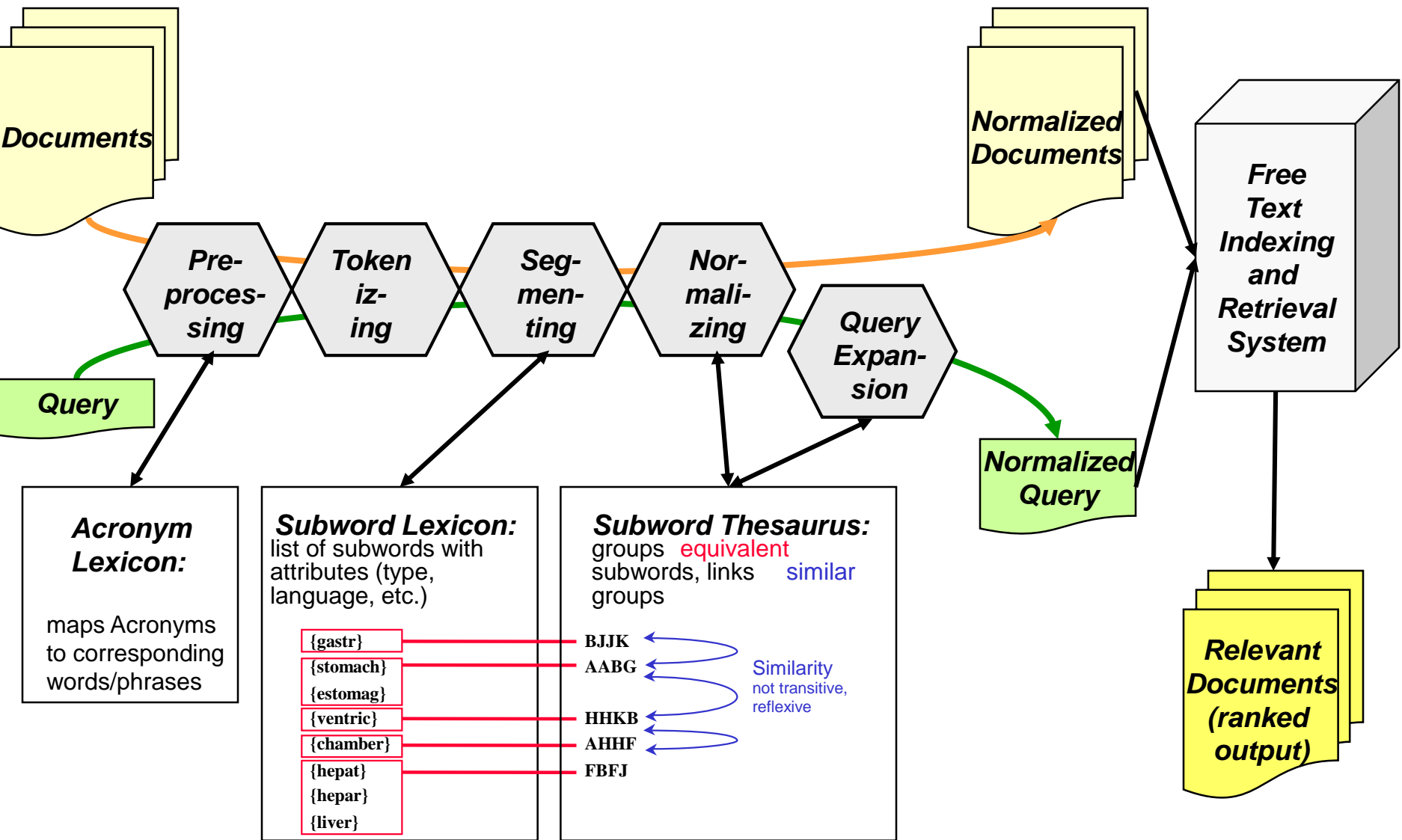
# Ergebnisse



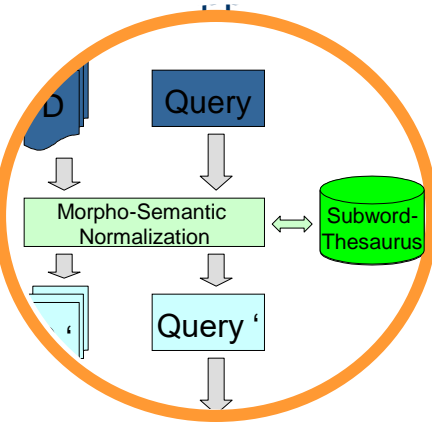
Precision-Recall-Diagramm:  
- Precisionwerte an fixen Recall-Leveln durch Interpolation  
- Mittelwert aus 25 Messreihen

1. Wortindex
2. Wortindex mit Stammformenreduktion
3. Subwordindex ohne Semantik
4. Subwordindex mit Semantik

# Extended System Architecture



# Lexical Resources



**Subword Lexicon:**  
list of subwords with attributes (type, language, etc.)

- {gastr}
- {stomach}
- {magen}
- {ventric}
- {chamber}
- {hepat}
- {hepar}
- {liver}
- {kidney}
- {ren}
- {nier}

Equivalence  
transitive  
and reflexive

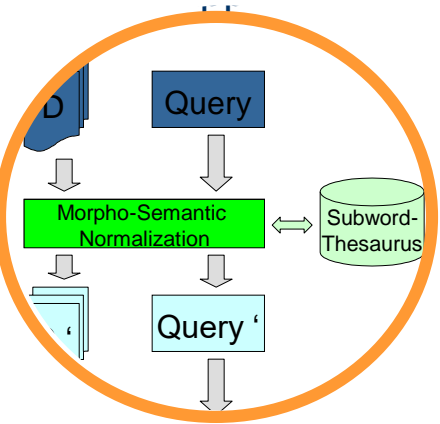
**Subword Thesaurus:**  
groups **equivalent** subwords,  
links similar groups

ID#

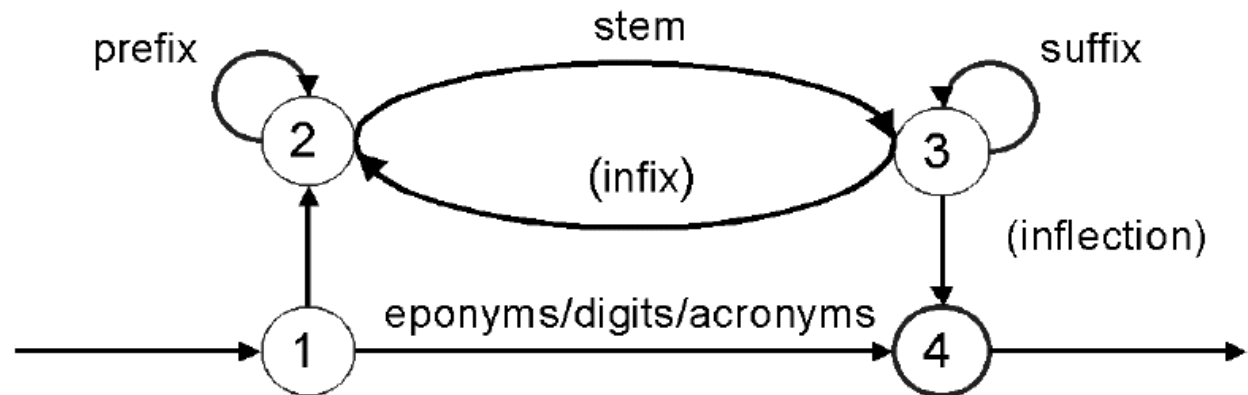
- \$5223\$
- \$6776\$
- \$3401\$
- \$7445\$
- \$9004\$
- \$6761\$

Similarity  
not transitive,  
reflexive

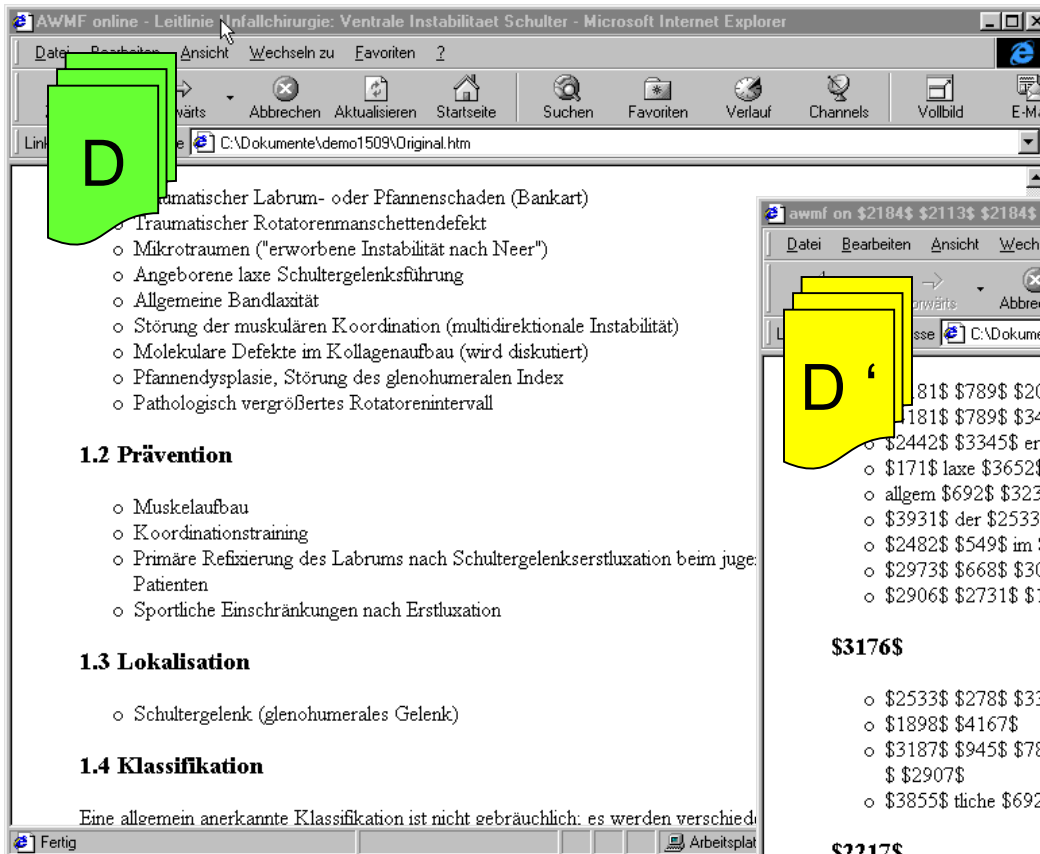
# Algorithmic Resources



- Morphosyntactic parser based on a word model described as a finite-state automaton
- Heuristic rules for disambiguation of parses



# Morphosemantic Normalization



AWMF online - Leitlinie Unfallchirurgie: Ventrale Instabilität Schulter - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Wechseln zu Favoriten ?

Link

C:\Dokumente\demo1509\Original.htm

traumatischer Labrum- oder Pfannenschaden (Bankart)

- o Traumatischer Rotatorenmanschettendefekt
- o Mikrotraumen ("erworbene Instabilität nach Neer")
- o Angeborene laxe Schultergelenksführung
- o Allgemeine Bandlaxität
- o Störung der muskulären Koordination (multidirektionale Instabilität)
- o Molekulare Defekte im Kollagenaufbau (wird diskutiert)
- o Pfannendysplasie, Störung des glenohumeralen Index
- o Pathologisch vergrößertes Rotatorenintervall

**1.2 Prävention**

- o Muskelaufbau
- o Koordinationstraining
- o Primäre Refixierung des Labrums nach Schultergelenkserstluxation beim jugendlichen Patienten
- o Sportliche Einschränkungen nach Erstluxation

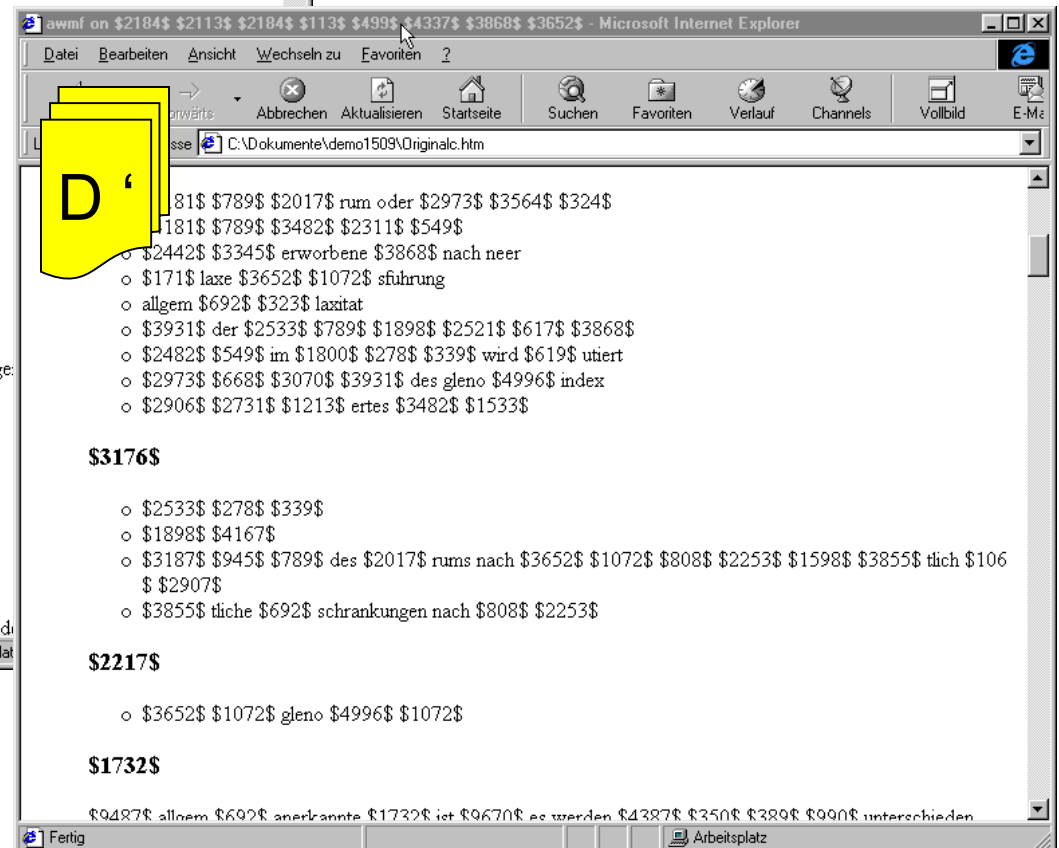
**1.3 Lokalisation**

- o Schultergelenk (glenohumerales Gelenk)

**1.4 Klassifikation**

Eine allgemein anerkannte Klassifikation ist nicht gebräuchlich: es werden verschiedene

Fertig Arbeitsplatz



awmf on \$2184\$ \$2113\$ \$2184\$ \$113\$ \$499\$ \$4337\$ \$3868\$ \$3652\$ - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Wechseln zu Favoriten ?

Link

C:\Dokumente\demo1509\Original.htm

81\$ \$789\$ \$2017\$ rum oder \$2973\$ \$3564\$ \$324\$

181\$ \$789\$ \$3482\$ \$2311\$ \$549\$

- o \$2442\$ \$3345\$ erworbene \$3868\$ nach neer
- o \$171\$ laxe \$3652\$ \$1072\$ sführung
- o allgem \$692\$ \$323\$ laxitat
- o \$3931\$ der \$2533\$ \$789\$ \$1898\$ \$2521\$ \$617\$ \$3868\$
- o \$2482\$ \$549\$ im \$1800\$ \$278\$ \$339\$ wird \$619\$ utiert
- o \$2973\$ \$668\$ \$3070\$ \$3931\$ des gleno \$4996\$ index
- o \$2906\$ \$2731\$ \$1213\$ ertes \$3482\$ \$1533\$

**\$3176\$**

- o \$2533\$ \$278\$ \$339\$
- o \$1898\$ \$4167\$
- o \$3187\$ \$945\$ \$789\$ des \$2017\$ rums nach \$3652\$ \$1072\$ \$808\$ \$2253\$ \$1598\$ \$3855\$ tlich \$106\$ \$2907\$
- o \$3855\$ tliche \$692\$ schrankungen nach \$808\$ \$2253\$

**\$2217\$**

- o \$3652\$ \$1072\$ gleno \$4996\$ \$1072\$

**\$1732\$**

\$9487\$ allgem \$692\$ anerkannte \$1732\$ ist \$9670\$ es werden \$4387\$ \$350\$ \$389\$ \$990\$ unterschieden

Fertig Arbeitsplatz

# Crossing Languages in Text Retrieval via an Interlingua

**Udo Hahn**    Kornél Markó

Michael Poprat    Stefan Schulz

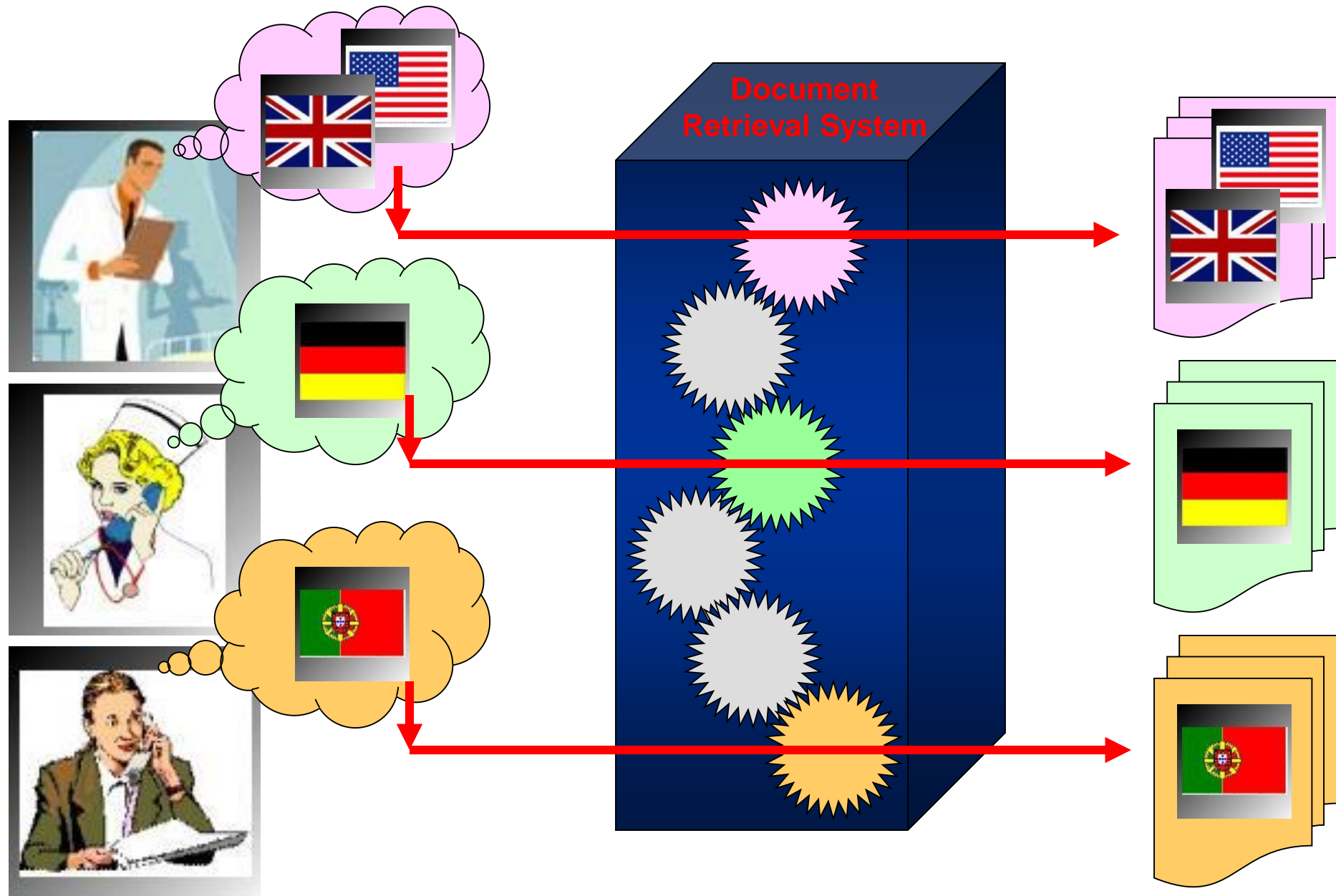
Joachim Wermter    Percy Nohama



Text Knowledge Engineering Lab  
Medical Informatics Division  
Freiburg University, Germany

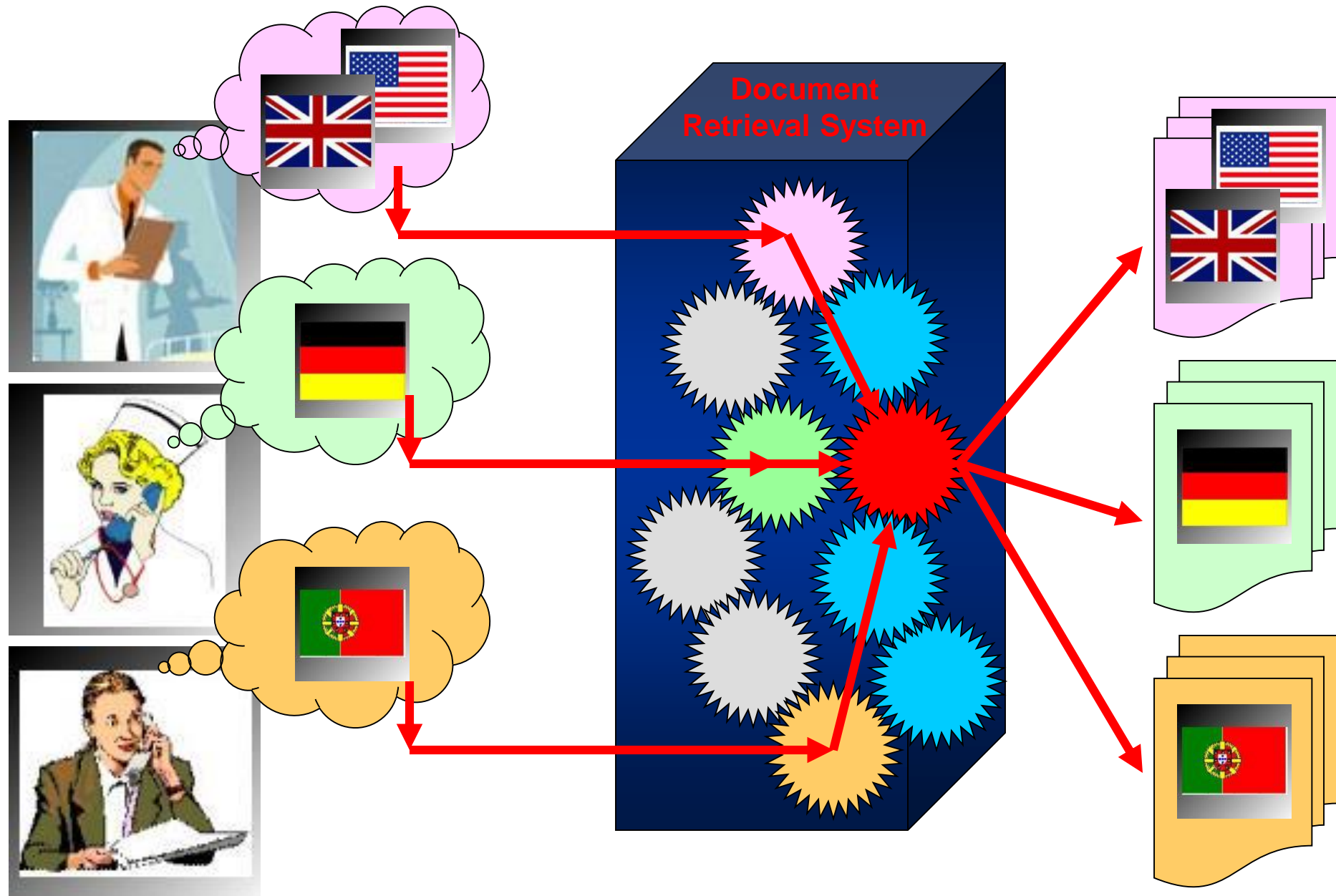
<http://www.coling.uni-freiburg.de>

# Monolingual Document Retrieval





# Crosslingual Document Retrieval



# Subword Lexicon & Thesaurus

## *Subword Lexicon:*

list of subwords

gastr  
stomach  
magen

ventric  
chamber

hepat, hepar  
liver  
leber

nephr  
ren  
kidney  
nier

Equivalence  
transitive  
and reflexive

## *Subword Thesaurus:*

grouping of near-synonymous  
subwords into **equivalence  
classes**

#GASTR

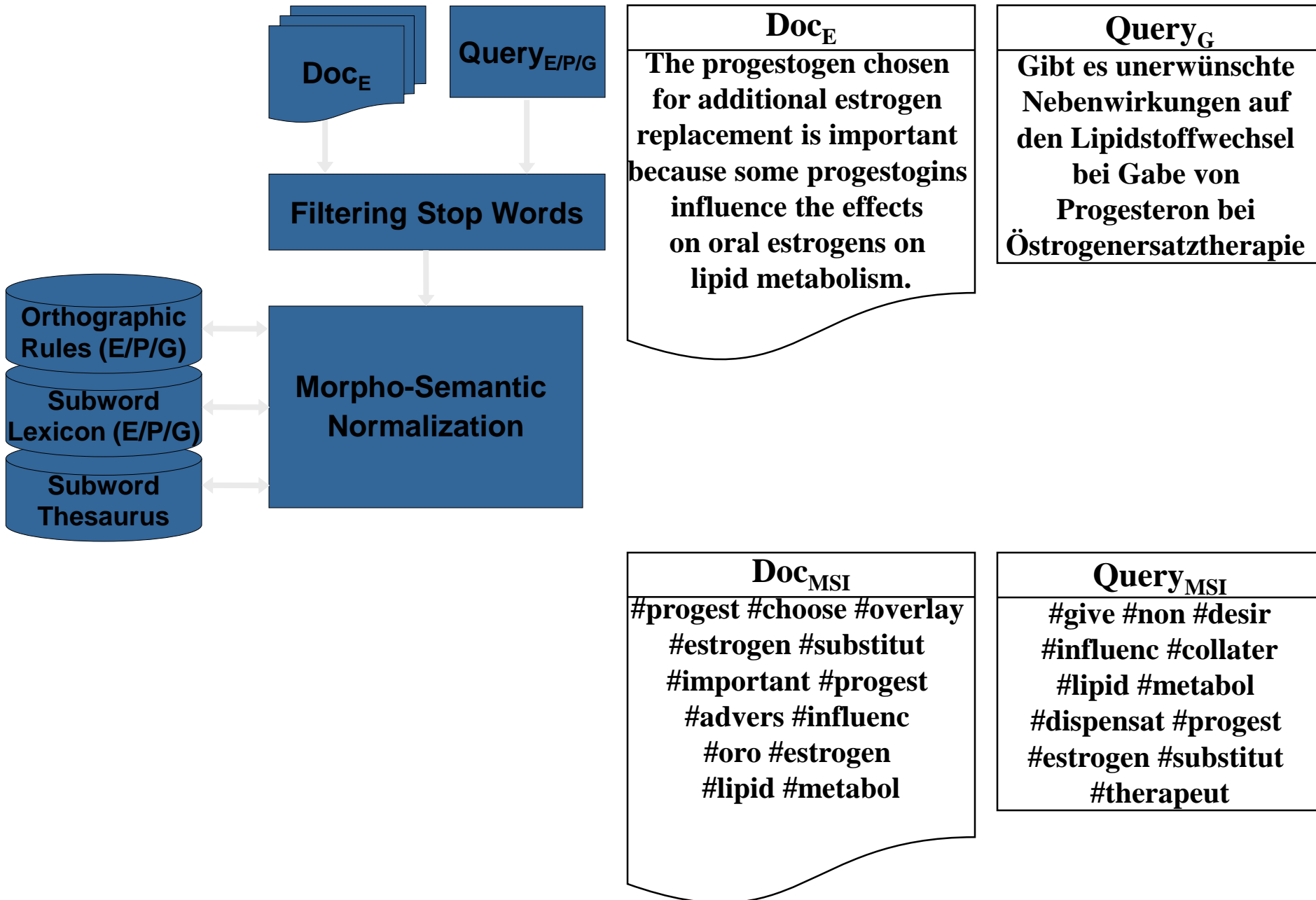
#CHAMBER

#HEPAR

#NEPHR

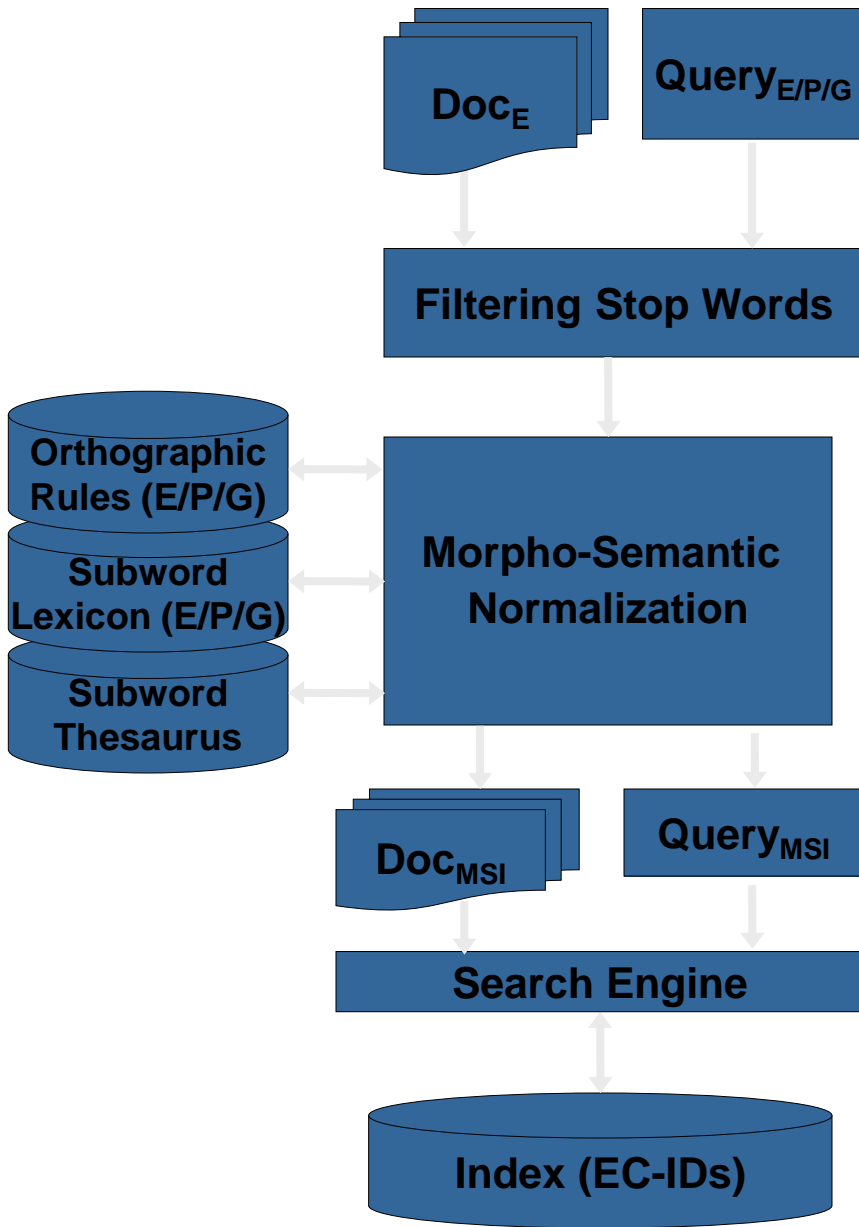
# Morpho-Semantic Indexing

# MorphoSaurus System



# Morpho-Semantic Indexing

# MorphoSaurus System



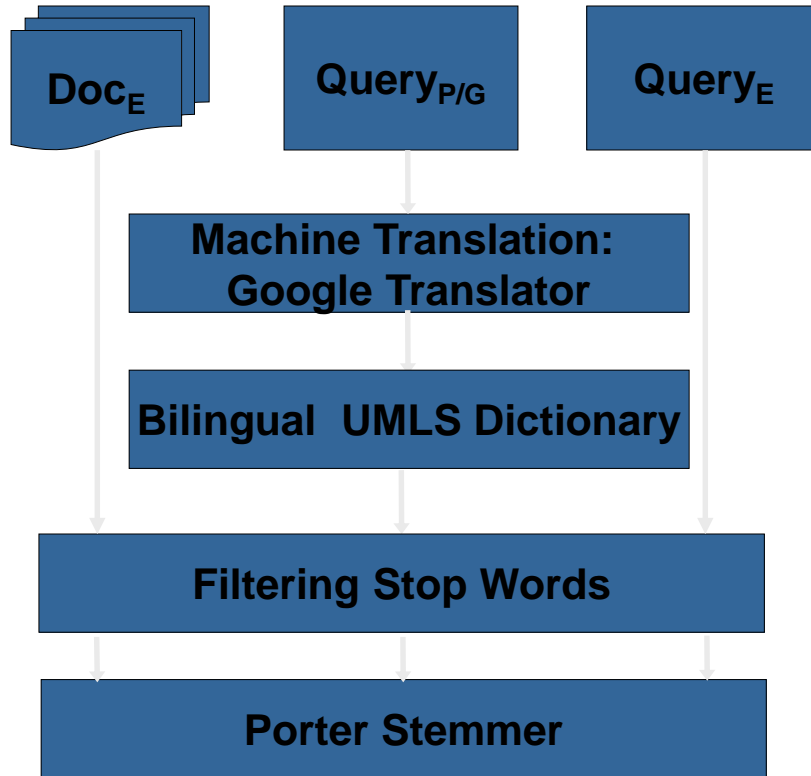
**Doc<sub>E</sub>**  
The progestogen chosen for additional estrogen replacement is important because some progestogins influence the effects on oral estrogens on lipid metabolism.

**Query<sub>G</sub>**  
Gibt es unerwünschte Nebenwirkungen auf den Lipidstoffwechsel bei Gabe von Progesteron bei Östrogensatztherapie

**Doc<sub>MSI</sub>**  
#progest #choose #overlay  
#estrogen #substitut  
#important #progest  
#advers #influenc  
#oro #estrogen  
#lipid #metabol

**Query<sub>MSI</sub>**  
#give #non #desir  
#influenc #collater  
#lipid #metabol  
#dispensat #progest  
#estrogen #substitut  
#therapeut

# Direct Query Translation



**Doc<sub>E</sub>**

The progestogen chosen for additional estrogen replacement is important because some progestogins influence the effects on oral estrogens on lipid metabolism.

**Query<sub>G</sub>**

Gibt es unerwünschte Nebenwirkungen auf den Lipidstoffwechsel bei Gabe von Progesteron bei Östrogensatztherapie

**Translated Query<sub>G→E</sub>**

There are unwanted side effects on the **Lipidstoffwechsel** with gift of progesteron with **Östrogensatztherapie**

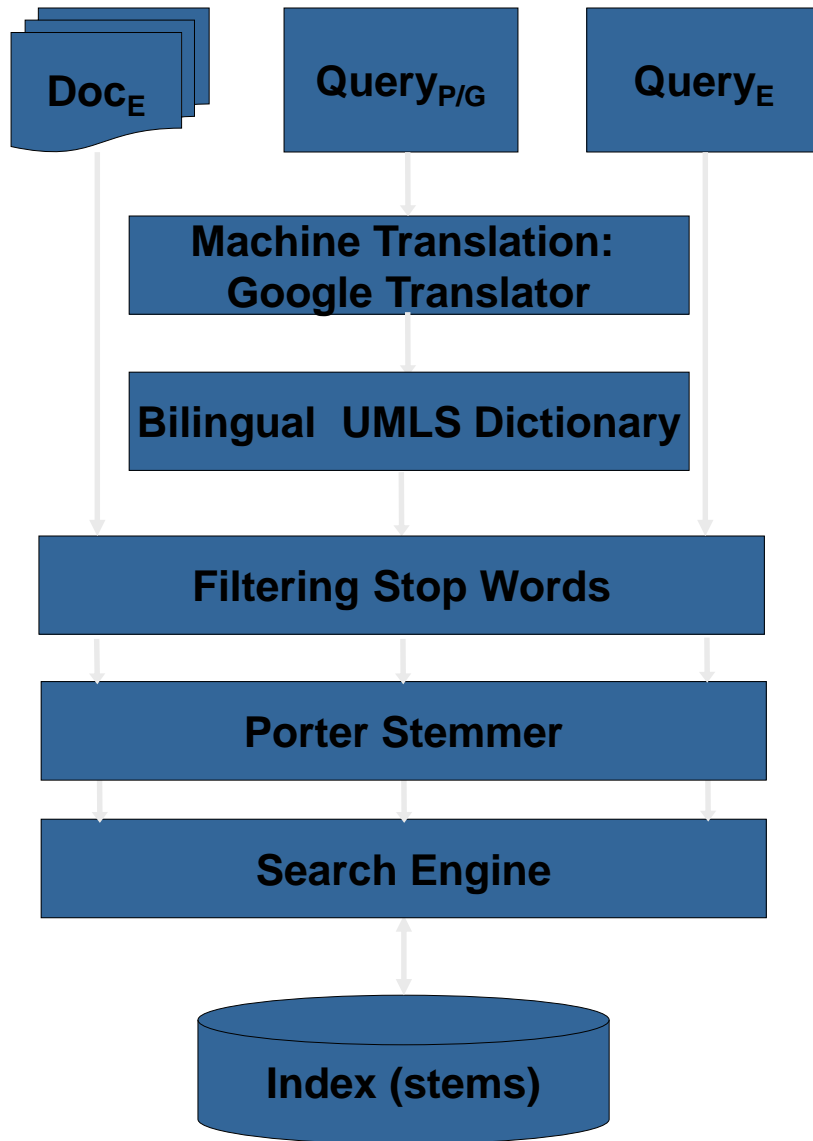
**Stemmed Doc<sub>E</sub>**

progestogen chosen addit  
estrogen replac import  
progestogin  
influenc effect oral  
estrogen lipid metabol

**Stemmed Query<sub>G→E</sub>**

unwant side effect  
**Lipidstoffwechsel**  
gift progesteron  
**Östrogensatztherapie**

# Direct Query Translation



**Doc<sub>E</sub>**  
The progestogen chosen for additional estrogen replacement is important because some progestogins influence the effects on oral estrogens on lipid metabolism.

**Query<sub>G</sub>**  
Gibt es unerwünschte Nebenwirkungen auf den Lipidstoffwechsel bei Gabe von Progesteron bei Östrogensatztherapie

**Translated Query<sub>G→E</sub>**  
There are unwanted side effects on the **Lipidstoffwechsel** with gift of progesteron with **Östrogensatztherapie**

**Stemmed Doc<sub>E</sub>**  
progestogen chosen addit  
estrogen replac import  
progestogin  
influenc **effect** oral  
estrogen lipid metabol

**Stemmed Query<sub>G→E</sub>**  
unwant side **effect**  
**Lipidstoffwechsel**  
gift progesteron  
**Östrogensatztherapie**

## Original Document / Query

### Doc<sub>E</sub>

The progestogen chosen for additional estrogen replacement is important because some progestogens influence the effects on oral estrogens on lipid metabolism.

### Query<sub>G</sub>

Gibt es unerwünschte Nebenwirkungen auf den Lipidstoffwechsel bei Gabe von Progesteron bei Östrogensatztherapie

## Direct Query Translation (QTR)

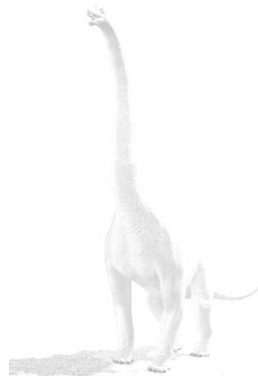
### Stemmed Doc<sub>E</sub>

progestogen chosen addit  
estrogen replac import  
progestogin  
influenc **effect** oral  
estrogen lipid metabol

### Stemmed Query<sub>G→E</sub>

unwant side **effect**  
Lipidstoffwechsel  
gift progesteron  
Östrogensatztherapie

## Morpho-Semantic Indexing (MSI)



### Doc<sub>MSI</sub>

**#progest** #choose #overlay  
**#estrogen** #substitut  
#important **#progest**  
#advers **#influnc**  
#oro **#estrogen**  
**#lipid** **#metabol**

### Query<sub>MSI</sub>

#give #non #desir  
**#influnc** #collater  
**#lipid** **#metabol**  
#dispensat **#progest**  
**#estrogen** **#substitut**  
#therapeut

# Experimental Setup

## ■ OHSUMED Corpus

- subset of the MEDLINE bibliographic database
- ~233,000 English documents (w./ abstracts only)
- 106 English queries



# Experimental Setup

## ■ OHSUMED Corpus

- subset of the MEDLINE bibliographic database
- ~233,000 English documents (w./ abstracts only)
- 106 English queries

## ■ **Subword Lexicons** (~58,000 entries, combined)

- English and German (~22,000 entries, each)
- Portuguese (~15,000 entries)

# Experimental Setup

## ■ OHSUMED Corpus

- subset of the MEDLINE bibliographic database
- ~233,000 English documents (w./ abstracts only)
- 106 English queries

## ■ Subword Lexicons (~58,000 entries, combined)

- English and German (~22,000 entries, each)
- Portuguese (~15,000 entries)

## ■ Subword Thesaurus

- ~22,000 equivalence classes

# Experimental Setup

## ■ OHSUMED Corpus

- subset of the MEDLINE bibliographic database
- ~233,000 English documents (w./ abstracts only)
- 106 English queries

## ■ Subword Lexicons (~58,000 entries, combined)

- English and German (~22,000 entries, each)
- Portuguese (~15,000 entries)

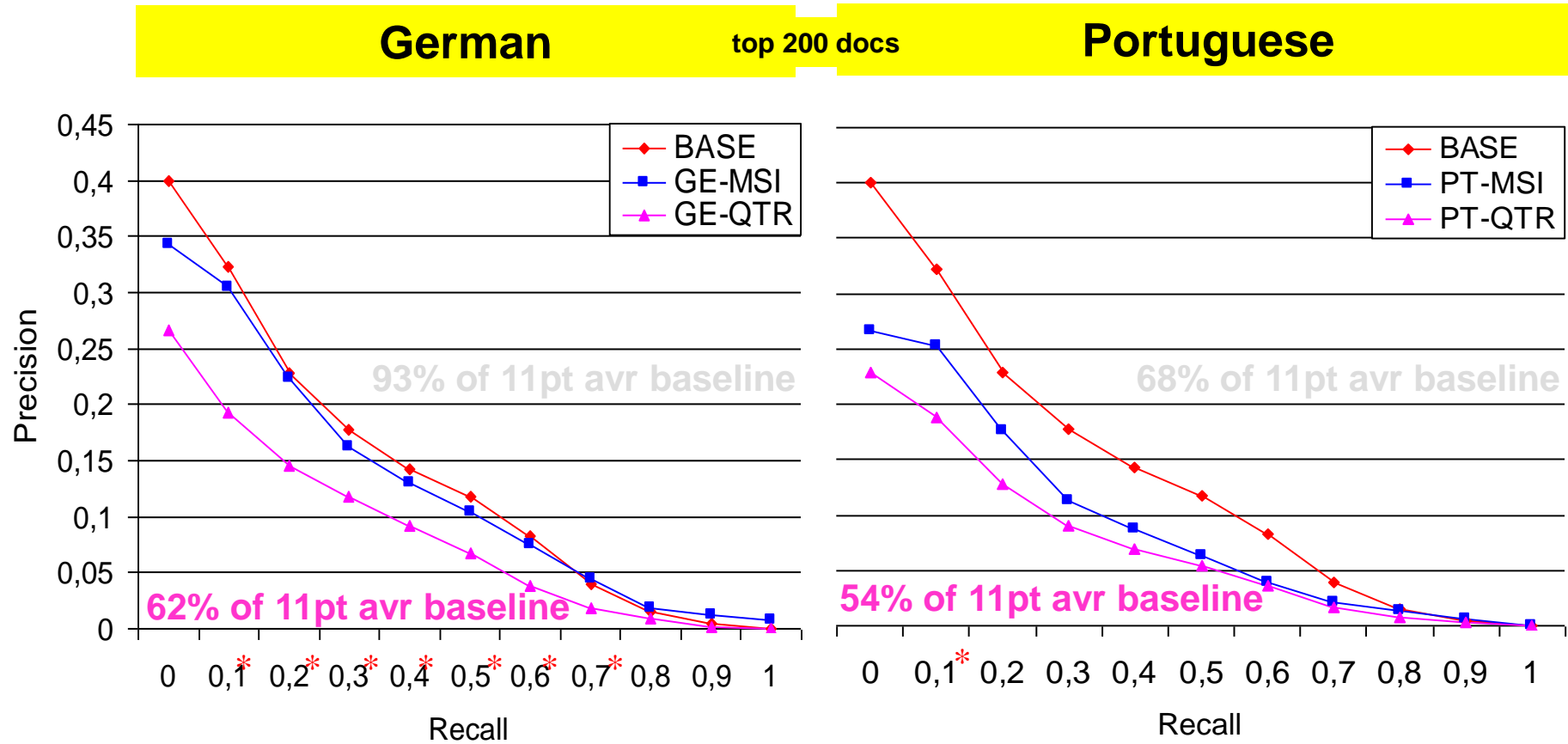
## ■ Subword Thesaurus

- ~22,000 equivalence classes

## ■ Test Conditions (Boolean search engine, ranked output)

- **BASE**: Porter-stemmed, stopped E docs & E queries
- **QTR**: GOOGLE & UMLS-translated, stopped G P queries
- **MSI**: morpho-semantically indexed G P queries

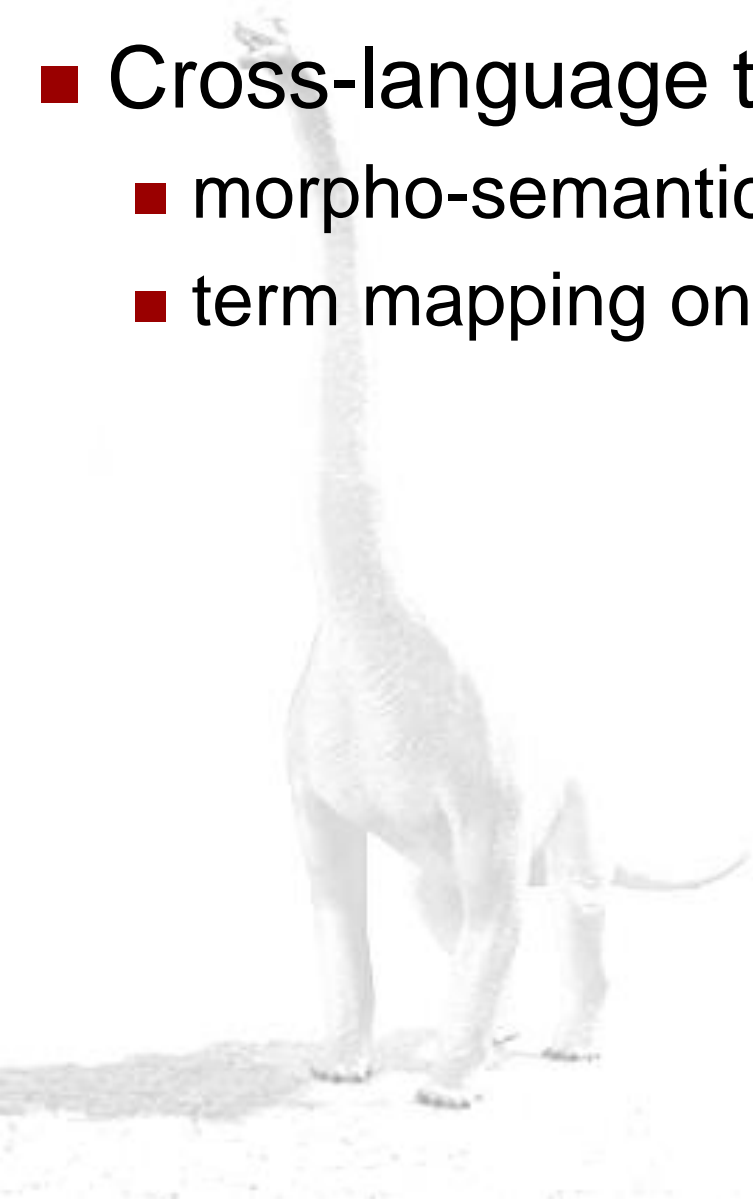
# Retrieval Performance



- BASE:** Porter stemming (E docs, E queries), stopped
- MSI:** Morpho-Semantic Indexing (G P docs, G P queries)
- QTR:** GOOGLE & UMLS translation of G P queries, Porter stemming (E docs, E queries), stopped

# Conclusions

- Cross-language text retrieval based on
  - morpho-semantic segmentation of docs & queries
  - term mapping on language-independent interlingua



# Conclusions

- Cross-language text retrieval based on
  - morpho-semantic segmentation of docs & queries
  - term mapping on language-independent interlingua
- Morpho-semantic indexing
  - achieves 93% of English baseline on German data (and 68% on Portuguese data)
  - outperforms direct query translation significantly
  - is independent from particular retrieval models

# Conclusions

- Cross-language text retrieval based on
  - morpho-semantic segmentation of docs & queries
  - term mapping on language-independent interlingua
- Morpho-semantic indexing
  - achieves 93% of English baseline on German data (and 68% on Portuguese data)
  - outperforms direct query translation significantly
  - is independent from particular retrieval models
- *MorphoSaurus* system runs on three languages: English, German, Portuguese

**<http://www.coling.uni-freiburg.de>**

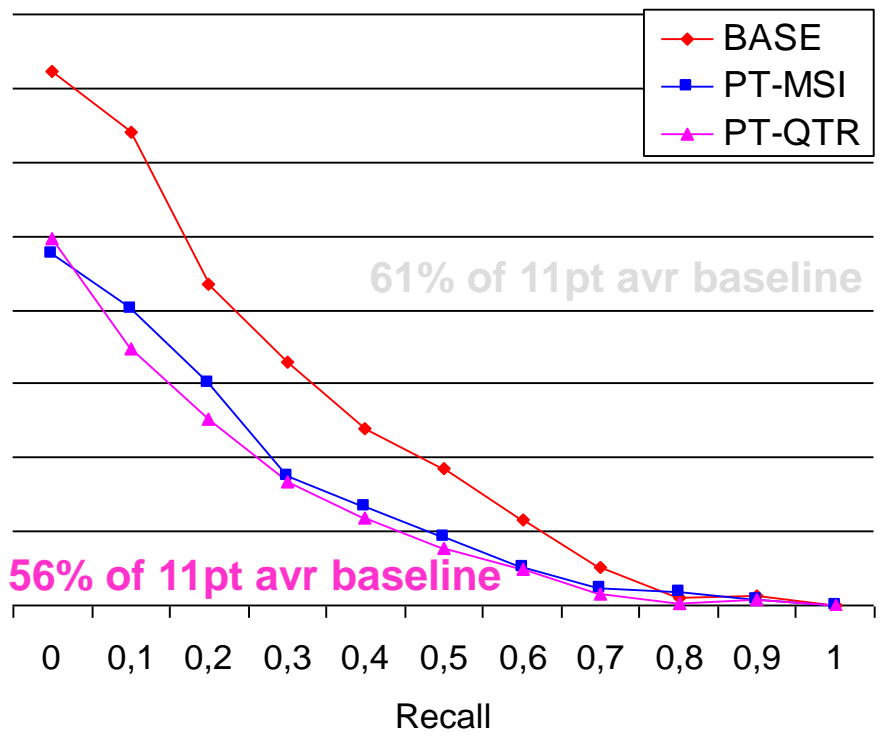
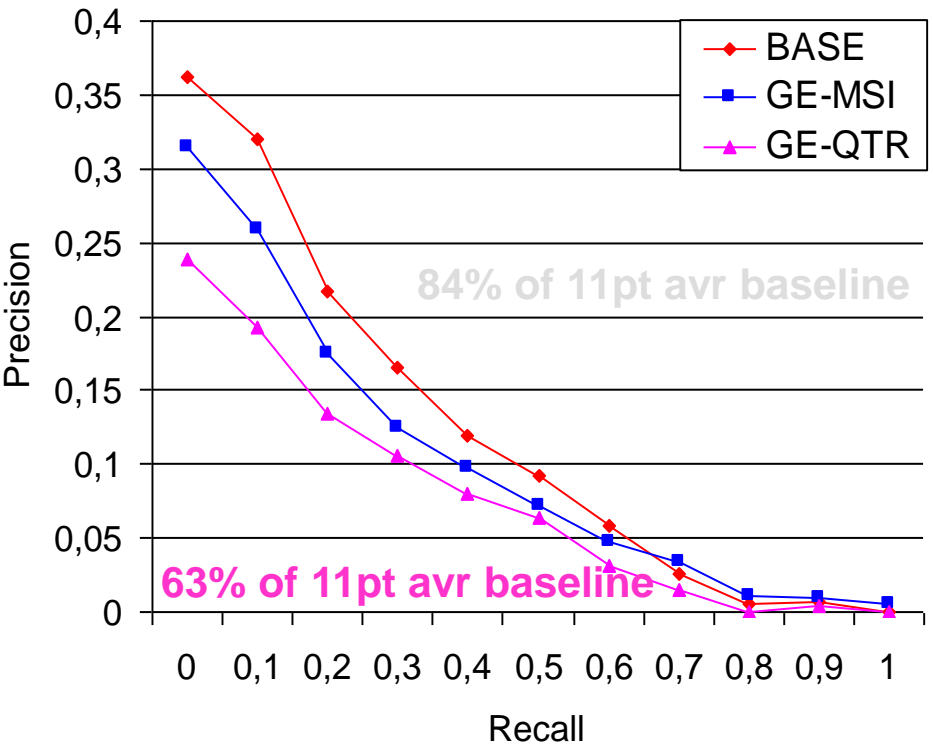


# Retrieval Performance without Adjacency

German

top 200 docs

Portuguese



- BASE:** Porter stemming (E docs, E queries), stopped
- MSI:** Morpho-Semantic Indexing (G P docs, G P queries)
- QTR:** GOOGLE & UMLS translation of G P queries, Porter stemming (E docs, E queries), stopped

# Text Retrieval Based on Medical Subwords

**Martin Honeck<sup>1</sup>, Udo Hahn<sup>2</sup>, Rüdiger Klar<sup>1</sup>, Stefan Schulz<sup>1</sup>**

<sup>1</sup> Department of Medical Informatics  
University Hospital Freiburg, Germany

<sup>2</sup> Natural Language Processing Division,  
Freiburg University, Germany

# Problem:

Poor performance of medical text retrieval in morphologically rich languages\*

\*most languages other than English

# Linguistic Phenomena hamper Medical Text Retrieval

## ■ Word formation

(inflection, derivation, composition):

*ulcus, ulcera, diagnosis, diagnoses, diagnostic, hepar, hepatic, para|sympath|ectomy, proct|o|sigmoid|o|scop|ie, Rechts|herz|insuffizienz*

## ■ Synonymy, spelling variants

*{oesophagus, esophagus}, {leuko, leuco}, {Magenulcus, Magenulkus}, {cutis, skin}, {hemorrhage, bleeding}, {ascorbic, Vitamin C}, {ancylostoma, hookworm}*

## ■ Multiple meanings:

*Cold {low temperature, common cold}, Bruch {fracture, hernia}, APA {antiperoxidase antibodies, american psychology association}*

# Example

- Frequency of German Word forms in *Google* Searches

Spelling Variants Synonyms			Inflections			Derivations		
Kolonkarzinom	2070	1780	Kolonkarzinom	2070	1770	Karzinom	17000	16900
Colonkarzinom	248	135	Kolonkarzinoms	471	253	karzinomatös	43	16
Coloncarcinom	111	73	Kolonkarzinome	275	139	karzinomatösen	86	40
Colon-Ca	203	169	Kolonkarzinomen	265	166	karzinomatöse	74	46
Kolon-Ca	66	46				karzinomatösem	7	5
Dickdarmkrebs	4000	3610				kazinomatöses	6	0
Dickdarmkarzinom	288	175				karzinomatöser	39	26
Dickdarmcarcinom	13	10						

**Number of Hits**

**Number of exclusive hits (no other form matches)**

# Hypothesis:

Improving Text Retrieval Performance using  
Linguistic Techniques

# Subword as Index Terms for Text Retrieval

- Subwords are atomic linguistic sense units :
  - Morphemes: *nephr, anti, thyr, scler, hepat, cardi*
  - Morpheme aggregates: *diaphys, ascorb, anabol, diagnost*
  - Words: *amyloid, bone, fever, liver*
  - (noun groups: *vitamin c,...*)
- Criterion: well-defined, non-decomposable medical concepts
- Grouping of synonymous subwords:
  - kkyxkj** = {*nephr, kidney, nier, ren*},
  - qxkjkq** = {*hepar, hepat, liver*},

# Resources

- Subword lexicons:  
Organize and classify subwords, prefixes and suffixes in several languages
- Subword thesaurus: Groups synonymous lexicon entries, links „similar“ groups
- Morphosyntactic parser: extracts subwords from text

Cf. Schulz et. al.

MEDINFO 2001

Yearbook of Medical Informatics '02



# Examples of Subword Extraction

## ■ Examples:

■ **proct** o **sigm** oid o **scop** y

■ **Schilddrüs** en **karzin** om

■ **cole cist ectom** ía

■ **acro cefal** o **sindattil** ia

■ **Sport verletz** ung en

■ **hør** sel s **hemm** ed e

■ **orchid** o **pex** ie

■ **Magen schleimhaut entzünd** ung

Lexical  
subwords  
(used for  
indexing)

Functional  
morphemes  
(not used for  
indexing)

# Experiment:

Does Subword-based medical text retrieval  
behave better than conventional methods ?

(formative evaluation - work in progress)

# Retrieval Experiments: Sources

- German version of the `Merck Manual` (medical textbook composed of 5,500 articles)
- 25 randomly chosen expert queries from medical students (German)
- 27 randomly chosen layman queries from the medical search engine “*Dr. Antonius*”
- Gold Standard:  
Three medical students did manual relevance assessment (52 \* 5,500 binary relevance judgements)

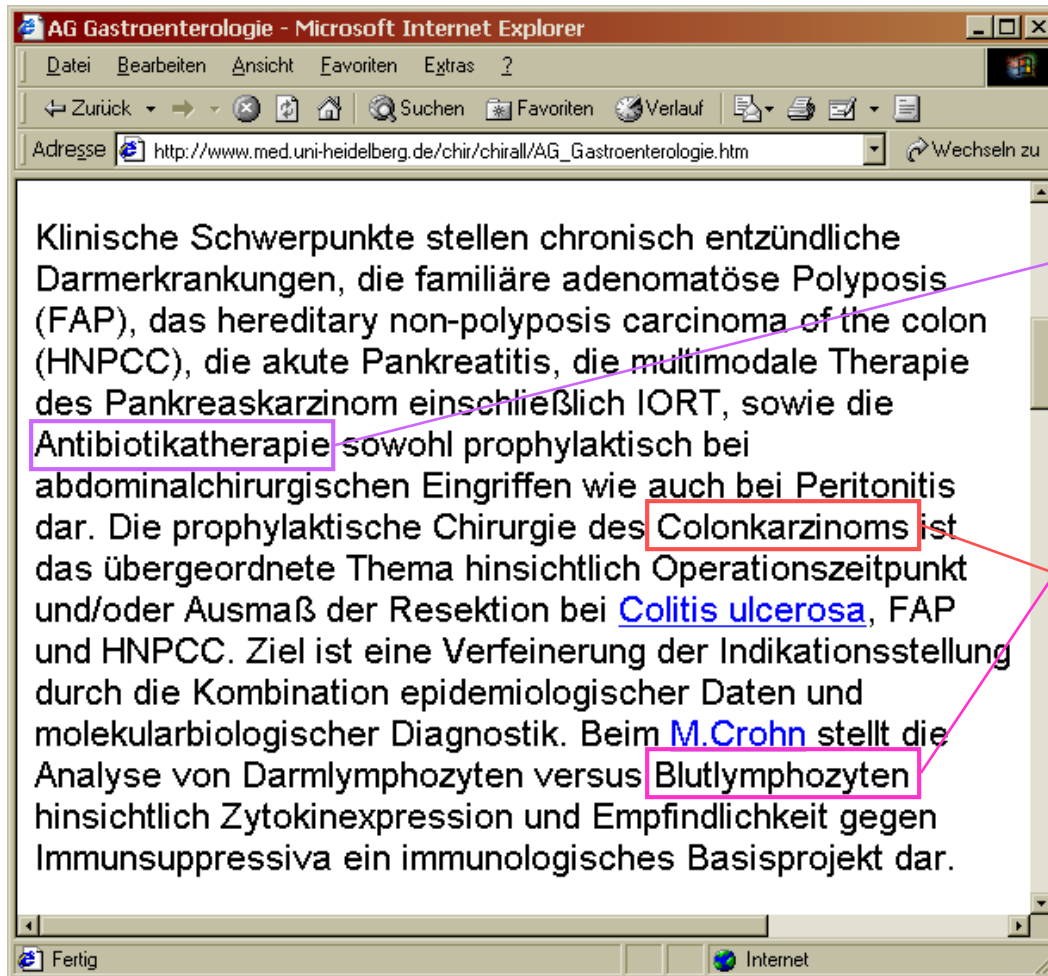
# Retrieval Experiments:

- Salton's Vector Space Retrieval Engine (produces ranked output)
- Proximity boost (proximity of query terms in documents matters for document ranking)
- Tests:
  - Test 1 (plain): Token Search. Baseline
  - Test 2 (segm): Morphological Segmentation
  - Test 3 (norm): Morphological Segmentation and Synonym Expansion.

For all tests:

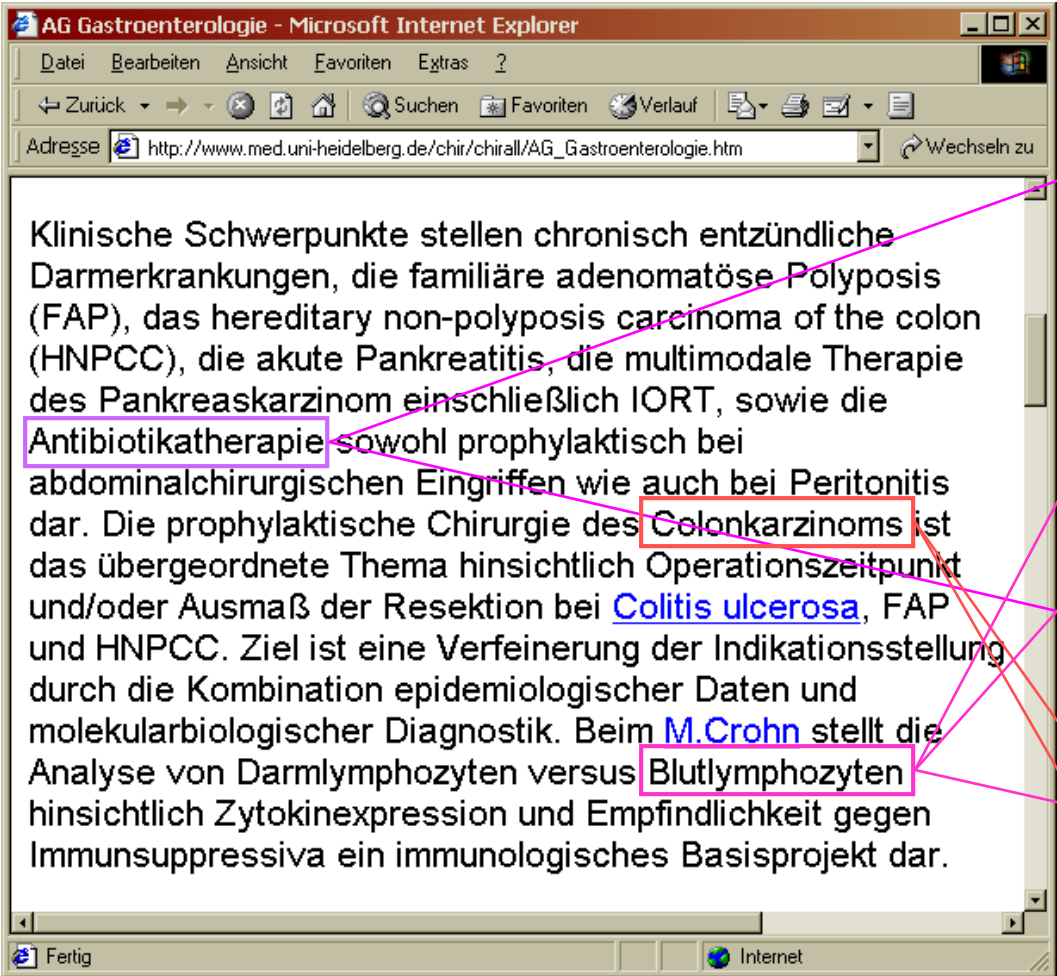
- Orthographic normalization preprocessing (e.g. ca → ka ,ci → zi, ä → ae, ...)

# Token-based Indexing



abdominalchirurgischen  
adenomatöse  
akute  
analyse  
antibiotikatherapie  
ausmaß  
basisprojekt  
blutlymphozyten  
carcinoma  
chirurgie  
chronisch  
colitis  
colon  
colonkarzinoms  
darmerkrankungen  
darmlymphozyten  
daten  
diagnostik  
eingriffen  
einschließlich  
empfindlichkeit  
entzündliche  
epidemiologischer

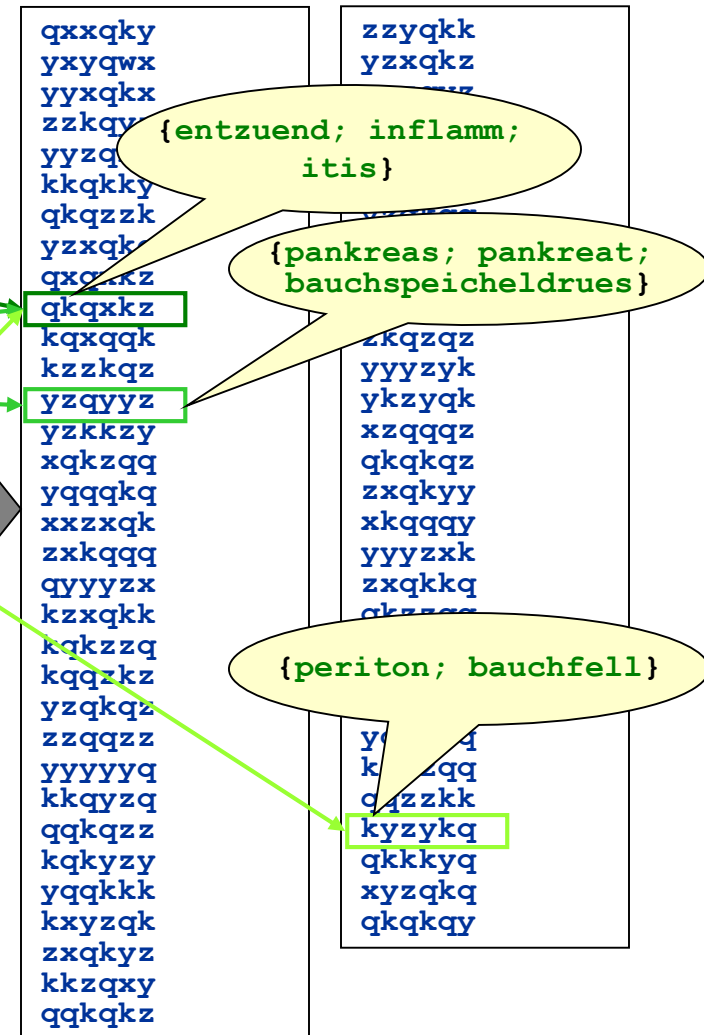
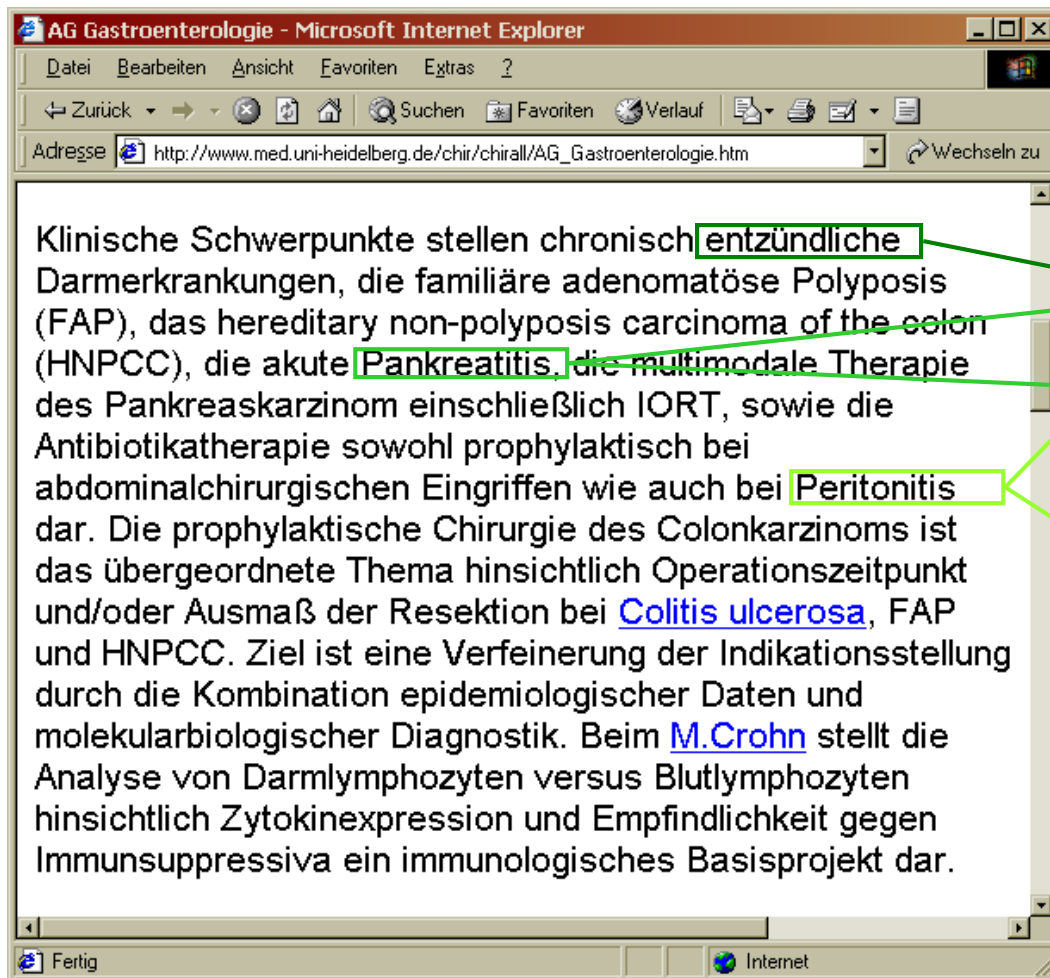
# Subword Indexing



- abdomin
- adenom
- akut
- analys
- antibiot
- ausmass
- basis
- biolog
- blut
- chirurg
- chroni
- darm
- daten
- diagnost
- eingriff
- empfindlich
- entzuend
- epidemiolog
- express
- famili
- fap
- fein
- heredit
- hinsichtlich
- hnpcc
- immun
- indik
- iort
- itis
- karzin
- klin
- kolitis
- kolon

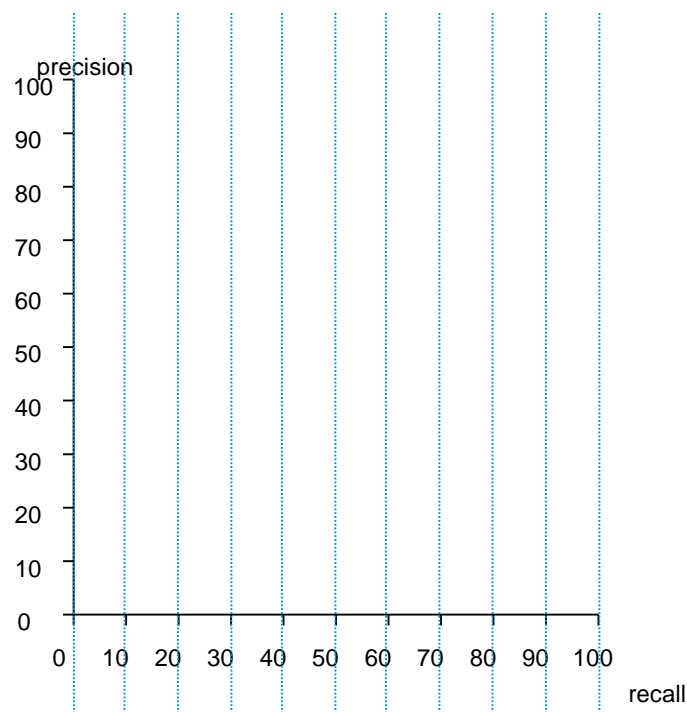
- kombin
- krank
- krohn
- lymph
- modal
- molekul
- multi
- non
- operation
- ordn
- osis
- pankreas
- pankreat
- periton
- polyp
- projekt
- prophylakt
- punkt
- resekt
- schwerpunkt
- stell
- suppress
- thema
- therap
- ueber
- ulzer
- versus
- zeit
- ziel
- zyt
- zytokin

# Subword - Indexing with Semantic Normalization



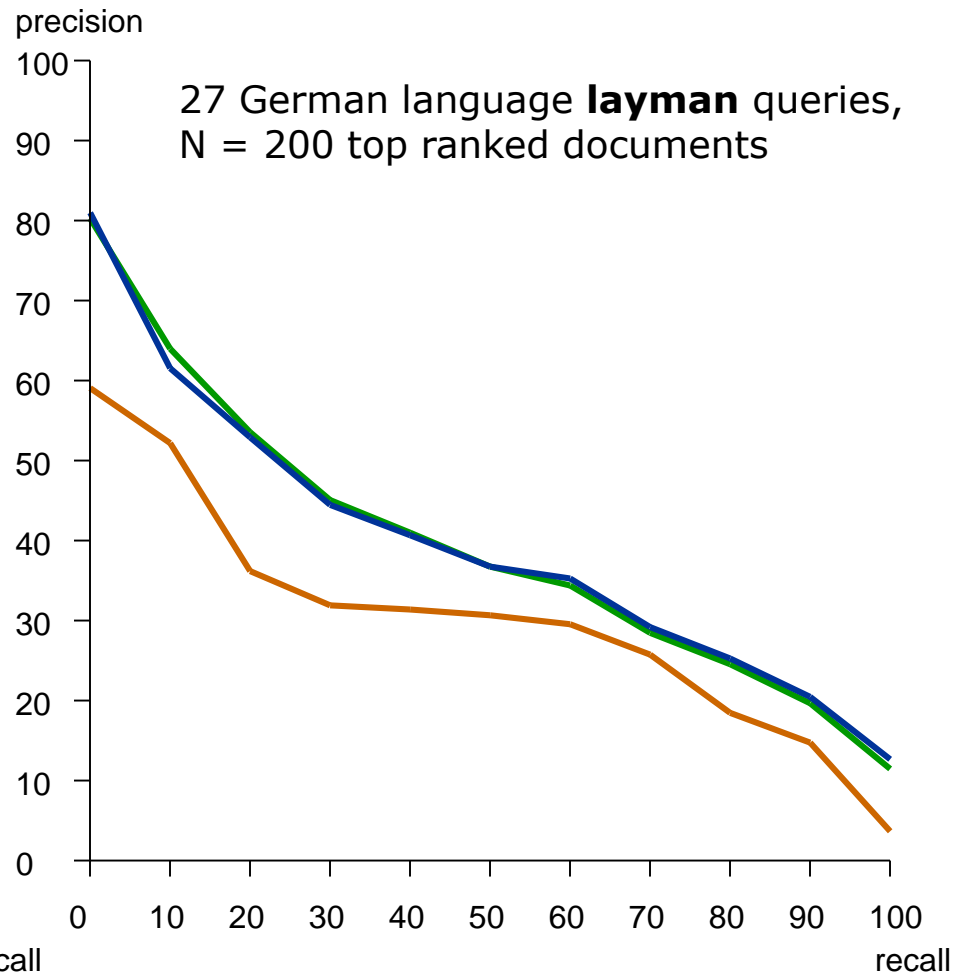
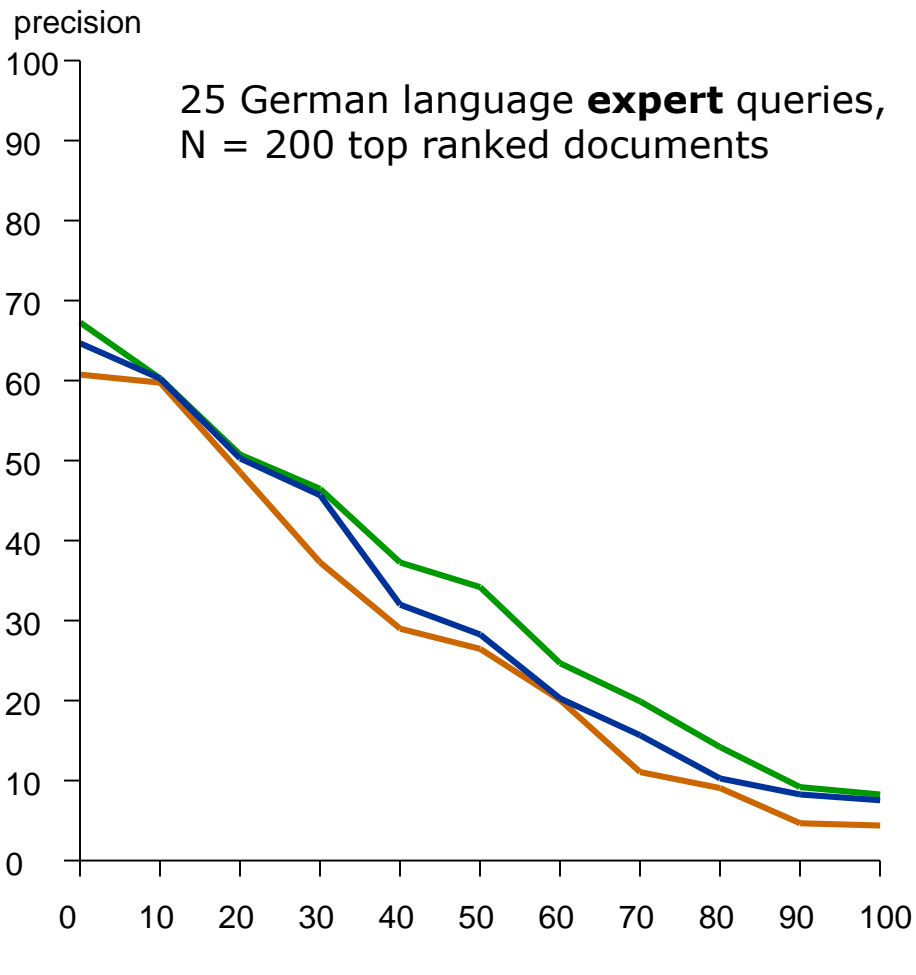
# Presentation of Results

- Precision / Recall Diagrams
- For each query:
  - interpolation of precision value at fixed recall levels (0%, 10%, ..., 100%)
- Arithmetic mean of precision values at each recall level





# Retrieval Experiments: Results



- **Test 1: Token Search ("plain"). Baseline**
- **Test 2: Morphological Segmentation ("segm")**
- **Test 3: Morphological Segmentation and Synonym Expansion. ("norm").**

# Significance Judgements

Precision (%)									
Recall (%)	<i>plain</i>	<i>segm</i>	<i>norm</i>	<i>plain</i>	<i>segm</i>	<i>norm</i>	<i>plain</i>	<i>segm</i>	<i>norm</i>
	expert queries n=25			layman queries n=27			all queries n=52		
0	60.8	67.3	64.7	59.1	<b>80.3</b>	<b>81.0</b>	60.0	<b>74.0</b>	<b>73.2</b>
10	59.8	60.3	60.3	52.2	64.0	61.6	55.8	62.3	61.0
20	48.6	50.8	50.3	36.2	<b>53.6</b>	<b>52.9</b>	42.1	52.3	51.7
30	37.3	46.5	45.7	31.9	<b>45.1</b>	<b>44.5</b>	34.5	<b>45.8</b>	<b>45.1</b>
40	29.0	37.3	32.0	31.4	<b>41.0</b>	<b>40.7</b>	30.3	39.2	36.5
50	26.5	34.2	28.3	30.7	36.8	36.8	28.7	<b>35.6</b>	32.7
60	20.1	24.7	20.3	29.6	34.4	35.3	25.0	29.7	28.1
70	11.1	19.9	15.7	25.8	28.5	29.2	18.7	24.4	22.7
80	9.1	14.2	10.3	18.5	24.6	25.3	14.0	<b>19.6</b>	18.1
90	4.7	<b>9.2</b>	8.3	14.8	19.7	<b>20.5</b>	9.9	<b>14.7</b>	<b>14.6</b>
100	4.4	<b>8.3</b>	7.6	3.7	<b>11.5</b>	<b>12.7</b>	4.0	<b>10.0</b>	<b>10.2</b>
11pt avrg.	24.1	33.9	31.2	29.5	40.0	40.0	26.9	37.0	35.8

$\alpha < 0.05$  (Wilcoxon test)

# Discussion:

Do the results justify the effort ?

# Discussion

- Work in progress
- Coverage of Subword dictionary (core vocabulary of clinical medicine (excl. proper names, acronyms) for German, English, Portuguese, ~ 17,000 entries). Target: 30,000 entries
- Linking subwords by synonymy relations adds noise to the system: more cautious use of synonymy relation
- Noise due to the erroneous extraction of medical subwords from non-medical terms and proper names: inclusion in dictionary

# Outlook

- Data-driven improvement of lexicons, thesaurus word grammar, algorithms, disambiguation heuristics
- Automated acquisition of abbreviations and acronyms (WWW)
- Semi-Automated acquisition of proper names
- Linkage to (MeSH): concept hierarchies, synonyms at the level of noun groups
- Evaluation of monolingual retrieval for Portuguese
- Evaluation of cross-lingual retrieval (German - English, English - Portuguese)

# Beispiel:

Klinische Schwerpunkte stellen chronisch entzündliche Darmerkrankungen, die familiäre adenomatöse Polyposis, die akute Pankreatitis, die multimodale Therapie des Pankreaskarzinoms, sowie die Antibiotikatherapie sowohl prophylaktisch als auch bei Peritonitis dar.

original

orthografische  
Normalisierung

klinische schwerpunkte stellen chronisch entzündliche darmerkrankungen die familiaere adenomatoese polyposis die akute pankreatitis die multimodale therapie des pankreaskarzinoms sowie die antibiotikatherapie sowohl prophylaktisch als auch bei peritonitis dar.

MorphoSaurus

Segmentierung

cliniijxqz	focusiipwxk	chronoiirjz
itidesiiixk	splanchniirqp	oticiiyyii
familiizxjr	adeniiwqz	oticiiyyii
polypiipjkw	oticiiyyii	acutaiijiz
pancreatiiqxir	itidesiiixk	multiikrkj
modaliiqxjr	therapiipri	pancreatiiqxir
oncoiijwqj	antibiosipypwr	therapiipri
prophylaktiipkiw	peritoniikzqx	itidesiiixk.

MID-Repräsentation

semantische  
Normalisierung

klin ische schwerpunkt e stell en chron isch entzünd liche darm erkrank ungen die famili aere adenom atoese polyp osis die akut e pankreat itis die multi modal e therap ie des pankreas karzinom s sowie die antibiotik a therap ie sowohl prophylakt isch als auch bei periton itis dar.



# Evaluation of Text Retrieval Systems

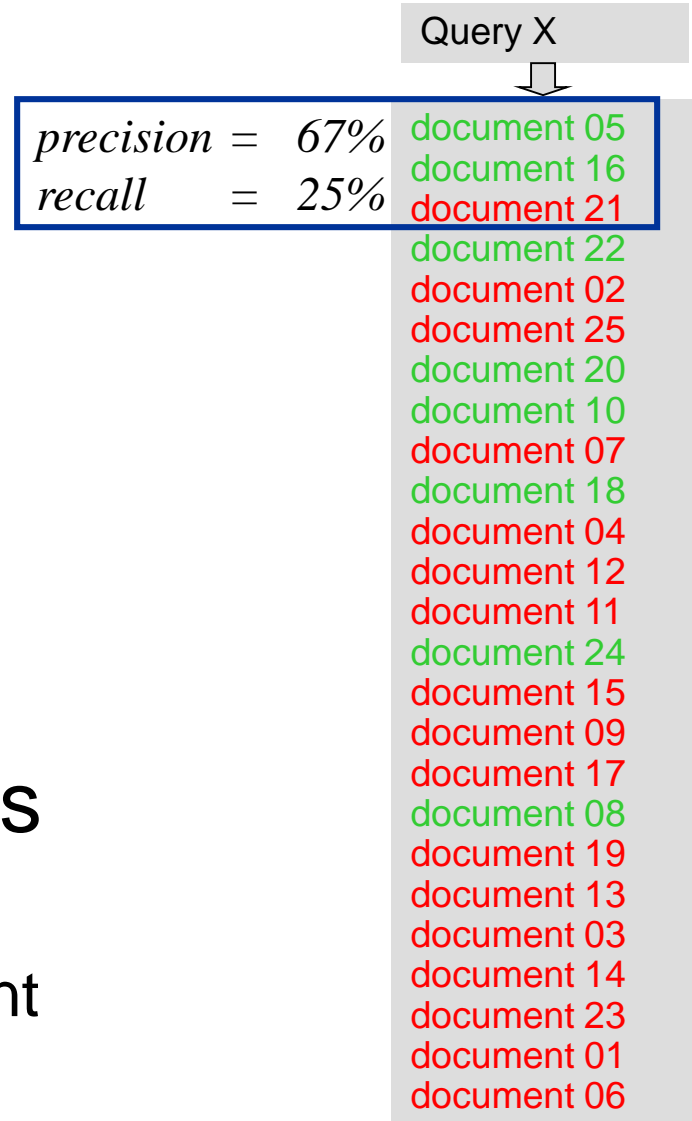
## ■ Target variables:

$$precision = \frac{n_{found+relevantDocuments}}{n_{found\_documents}}$$

$$recall = \frac{n_{found+relevant\_documents}}{n_{relevant\_documents}}$$

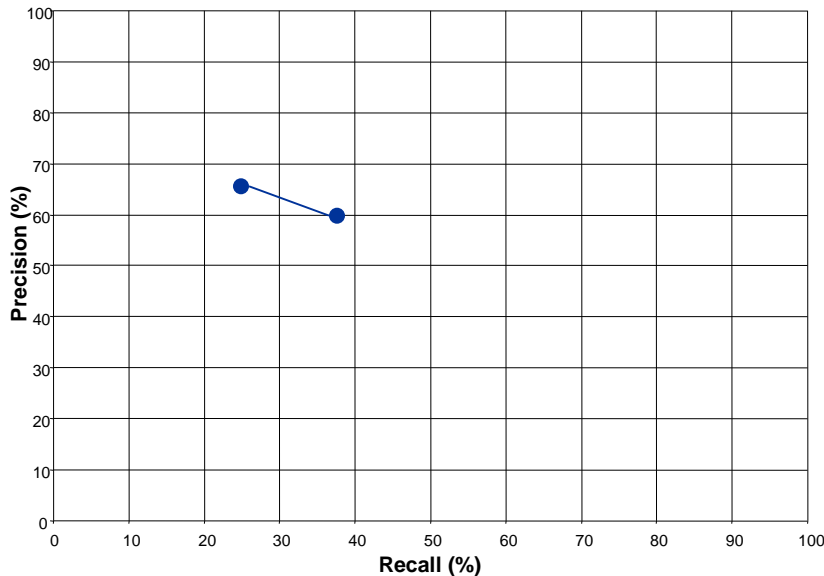
## ■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant





# Evaluation of Text Retrieval Systems



## ■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant

Query X

*precision* = 60%

*recall* = 38%

document 05  
document 16  
document 21  
document 22  
document 02  
document 25  
document 20  
document 10  
document 07  
document 18  
document 04  
document 12  
document 11  
document 24  
document 15  
document 09  
document 17  
document 08  
document 19  
document 13  
document 03  
document 14  
document 23  
document 01  
document 06

# Evaluation of Text Retrieval Systems

Query X

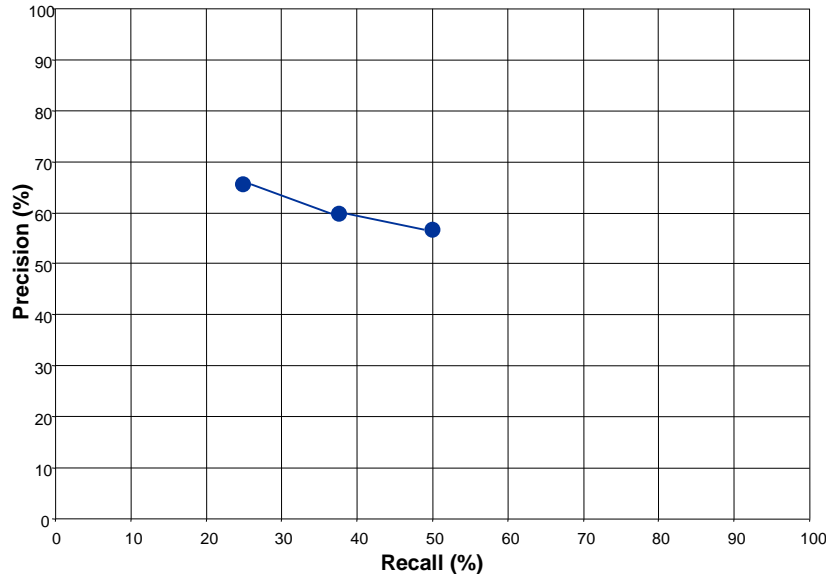


*precision* = 57%

*recall* = 50%

document 05  
document 16  
document 21  
document 22  
document 02  
document 25  
document 20

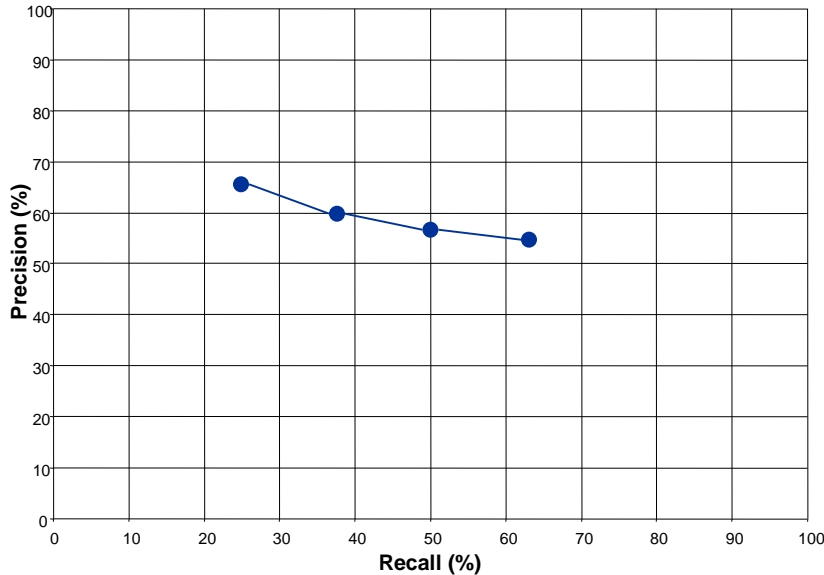
document 10  
document 07  
document 18  
document 04  
document 12  
document 11  
document 24  
document 15  
document 09  
document 17  
document 08  
document 19  
document 13  
document 03  
document 14  
document 23  
document 01  
document 06



## ■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant

# Evaluation of Text Retrieval Systems



## ■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant

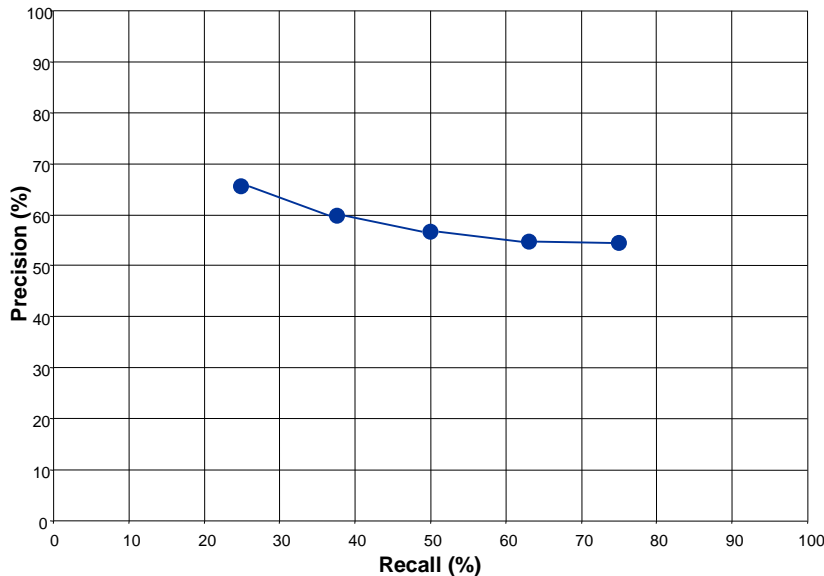
Query X

*precision* = 55%

*recall* = 63%

document 05  
document 16  
document 21  
document 22  
document 02  
document 25  
document 20  
document 10  
document 07  
document 18  
document 04  
document 12  
document 11  
document 24  
document 15  
document 09  
document 17  
document 08  
document 19  
document 13  
document 03  
document 14  
document 23  
document 01  
document 06

# Evaluation of Text Retrieval Systems



## ■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant

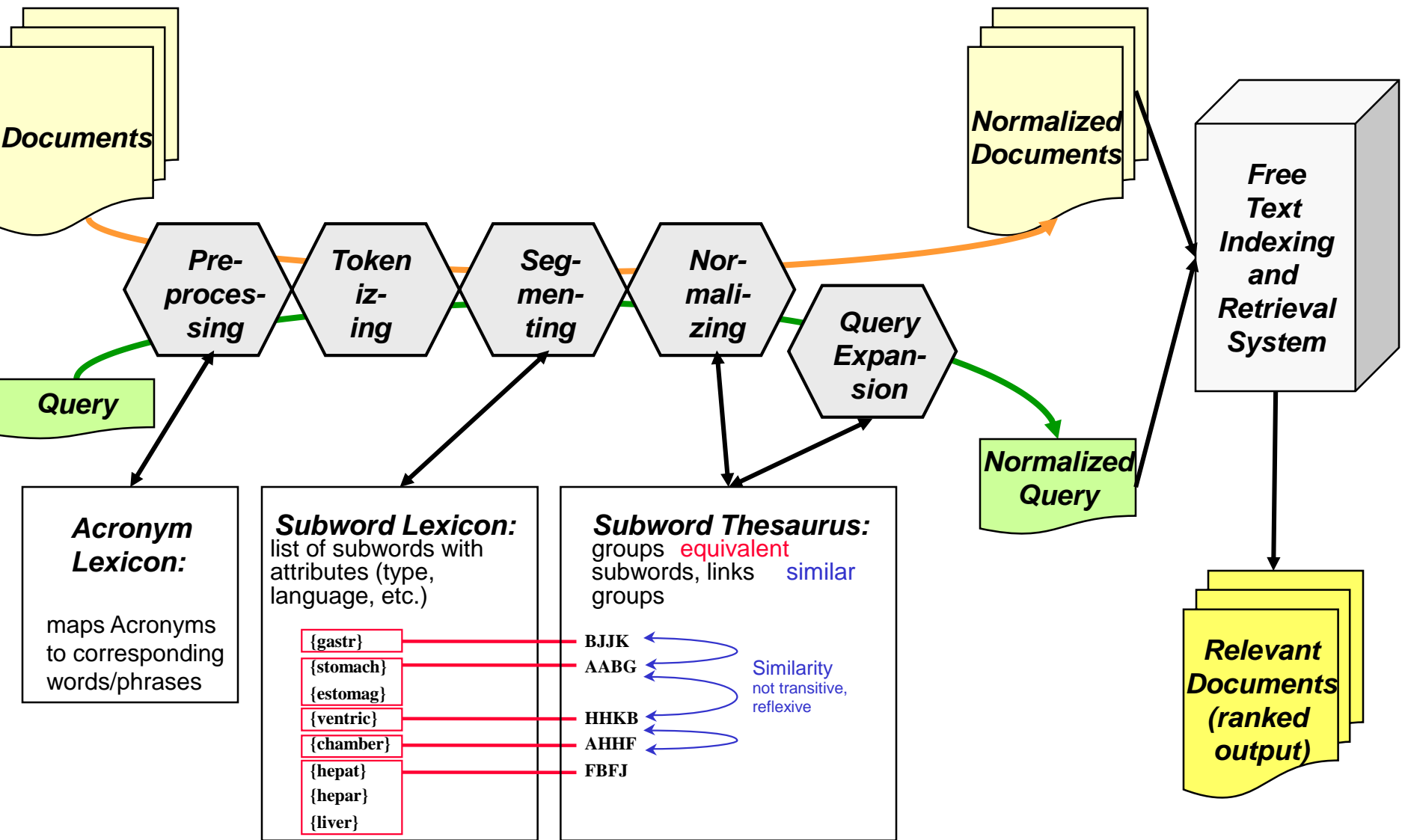
Query X

*precision* = 54%

*recall* = 75%

document 05  
document 16  
document 21  
document 22  
document 02  
document 25  
document 20  
document 10  
document 07  
document 18  
document 04  
document 12  
document 11  
document 24  
document 15  
document 09  
document 17  
document 08  
document 19  
document 13  
document 03  
document 14  
document 23  
document 01  
document 06

# Extended System Architecture



# Tool: Subword Editor & Workbench

editing tool for  
subword lexicon  
and thesaurus

**Segmentation**

Please enter text here:  
abdominal pain

Segmented text:  
ABDOMIN -al PAIN

Active Language  
 German  English

Db Update Update + Segment Segment Close

morphe thesaurus construction tool [C:\Dokumente\mck neu 008\thetemplate1909200...]

File Morphemes Thesaurus ?

Lexem A	Type	Weight	Lang
abdomen	1	3	1

a  
ab  
abdomen  
abdomin  
abduzens  
abduz  
abel  
aberr  
abfall  
abfluss  
abfuehr  
abgeleitet  
abhaeng

Find Match left truncation

Sort / Filter A

Lexem B	Type	Weight	Lang
belly	1	3	2

belly  
abdomin  
abduzens  
abdukt  
able  
aberr  
arthropod  
artificial  
like  
artikul  
arytaen  
aryten  
drug

Find Match left truncation

Sort / Filter B

4

abdomen  
abdomin  
unterleib

abdomen  
belly  
abdomin

5205

belly

5205

4

abdomen  
abdomin  
unterleib  
340  
bauch

Morphemes: 9462 EqClasses: 9064 SIM-Rel.: 6604 Compactness f(m/c) .73 Net Density f(r/c) 1.04

testbed for  
segmentation

# The Subword Approach (II)

- Language-specific algorithms for **extraction** of subwords from (medical) texts
- **Multilingual** subword **repositories**
- Criteria for subword delimitation and classification
  - Semantic (compositionality)  
*Hyper | cholesterol | emia*
  - Lexical (enabling synonym matching)  
*schleimhaut = mucosa (~~schleim | haut~~)*
  - Data-driven (avoiding ambiguities and false segmentation), e.g.  
*relation~~ship~~, Schwangers~~chaft~~ (~~relation | ship, Schwanger | schaft~~)*

# Disfunção tireoideana perinatal

As doenças da tireóide acometem 10% das mulheres, mas a maioria das pacientes responde bem ao tratamento.

Durante a gestação, mudanças metabólicas podem ocultar a patologia, com risco de dano fetal devido à conduta inapropriada. Exames de TSH, tiroxina livre e triiodotironina livre são essenciais.

Geralmente, a presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo. Este último costuma manifestar-se através de bócio, oftalmopatia, fraqueza muscular, taquicardia ou perda de peso.

# Perinatal Thyroid Dysfunction

Thyroid gland diseases affect 10% of women, but most patients respond well to treatment.

During pregnancy, metabolic changes can hide the presence of the disorder with the risk of fetal damage due to inappropriate handling. Measurement of "TSH", free "T4" and "T3" are indispensable.

Generally, high TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism. Typical manifestations of the latter are goiter, ophthalmopathy, muscular weakness, tachycardia, or weight loss.

**Original text (D)**

## DIS FUNCAO TIREOID e ana PERI NATAL

as DOENCA s da TIREOID e ACOMET em 10% das MULHER es MAS a MAIOR ia das PACIENT es RESPOND e BEM ao TRATAMENT o

DURANTE a GESTAC ao MUDANCA s METABOL ic as PRESENC a da PATOLOG ia COM RISC o de DAN a CONDU T a in APROPRIAD a. os EXAME s de "TSH", e TRI IODO TIRONIN a LIVR e sao ESSENCI ais

GERAL mente a PRESENC a de VALOR es ELEVAD os de "TSH" SUGER e o DIAGNOST ic o de HIPO TIREOID ism o PRIMAR io ENQUANTO NIVEIS SUPRIM id os de "TSH" SUGER em HIPER TIREOID ism o. este ULTIM o COSTUM a MANIFEST ar se ATRAVES de BOCIO, OFTALM o PATIA FRAQU eza MUSCUL ar TAQUI CARD ia ou PERD a de PESO.

## PERI NATAL THYROID DYS FUNCTION

THYROID GLAND DISEAS es AFFECT 10% of WOMEN BUT MOST PATIENT s RESPOND WELL to TREATMENT

DURING pregnancy METABOL ic CHANGE s CAN HIDE the DISORDER WITH the RISK of FETAL DAMAGE DUE to inAPPROPRIAD e HANDL ing. MEASURE ment of "TSH", FREE "T4" and "T3" are INDISPENSABLE

GENERAL ly HIGH "TSH" VALUE s SUGGEST the DIAGNOS is of PRIMAR y HYPO THYROID ism WHILE a SUPPRESS ed "TSH" LEVEL SUGGEST s HYPER THYROID ism. TYP ic al MANIFEST ation s of the LATTER are GOITER, OPHTALM o PATHY, MUSCUL ar WEAK ness TACHY CARD y or WEIGHT LOSS.

**Segmented text**

## iiiiill iiifunct iiithyr iiibirth

iiipatho iiithyr iiiaffect 10% iiifemin iiibut iiigh iiipatient iiirepond iiigood iiitreatment.

iiiduring iiipregnan iiichange iiimetabol iiipossibl iiiiiwith iiirisk iiidamage iiifetus iiidue iiibehav iiisuita iiihormon, iiithyroxin iiifree iiithree iiijod iiithyronin iiigeneral iiipresent iiivalue iiigh iiithyr iiistimul iiidiagnos iiilow iiithyr iiifirst iiiduring iiilevel iiisuppress iiithyr iiistimul iiihormon iiisuggest iiigh iiithyrii. iiilast iiicustom iiimanifest iiiby iiigoiter, iiieye iiipatho iiiiweak iiimucle iiispeed iiheart iiilose iiiveigh.

## iiibirth iiithyr iiiill iiifunct

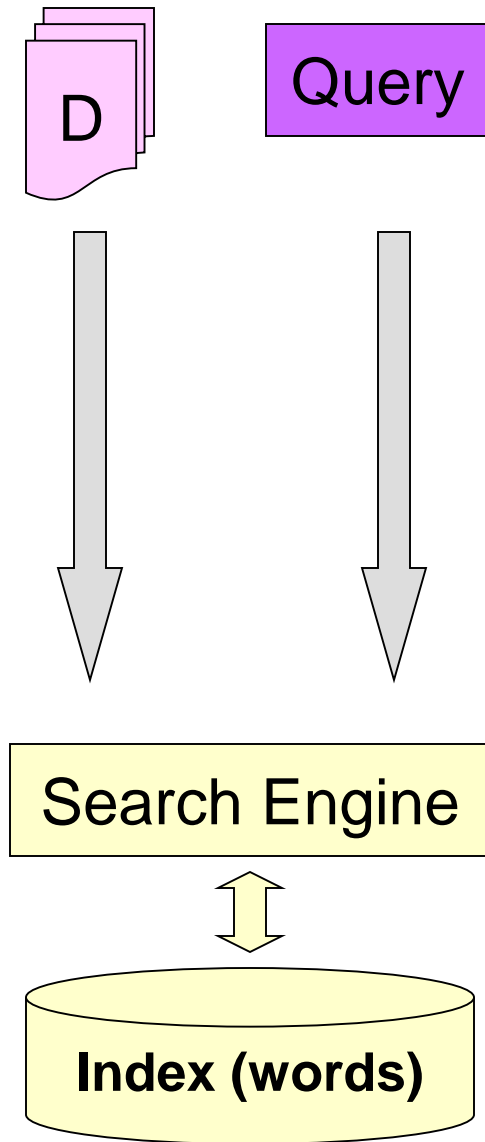
iiithyr iiigland iiipatho iiiaffect 10% iiifemin iiibut iiigh iiipatient iiirepond iiigood iiitreatment

ol iiichange iiican iihide iiipresent iiipatho iiiiwith iiie iiisuitabl iiimanag. iiimeasure iiithyr iiistimul iiithree iiijod iiithyronin iiieessential iiul iiihormon iiivalue iiisuggest iiidiagnos iiifirst iiilow iiithyr iiiduring iiisuppress iiithyr iiistimul iiihormon iiilevel iiisuggest iiigh iiithyr. iiityp iiimanifest iiilast iiigoiteriii, iiieye iiipathoiii, iiimucle iiiiweak iiispeed iiheart iiiveigh iiilose..

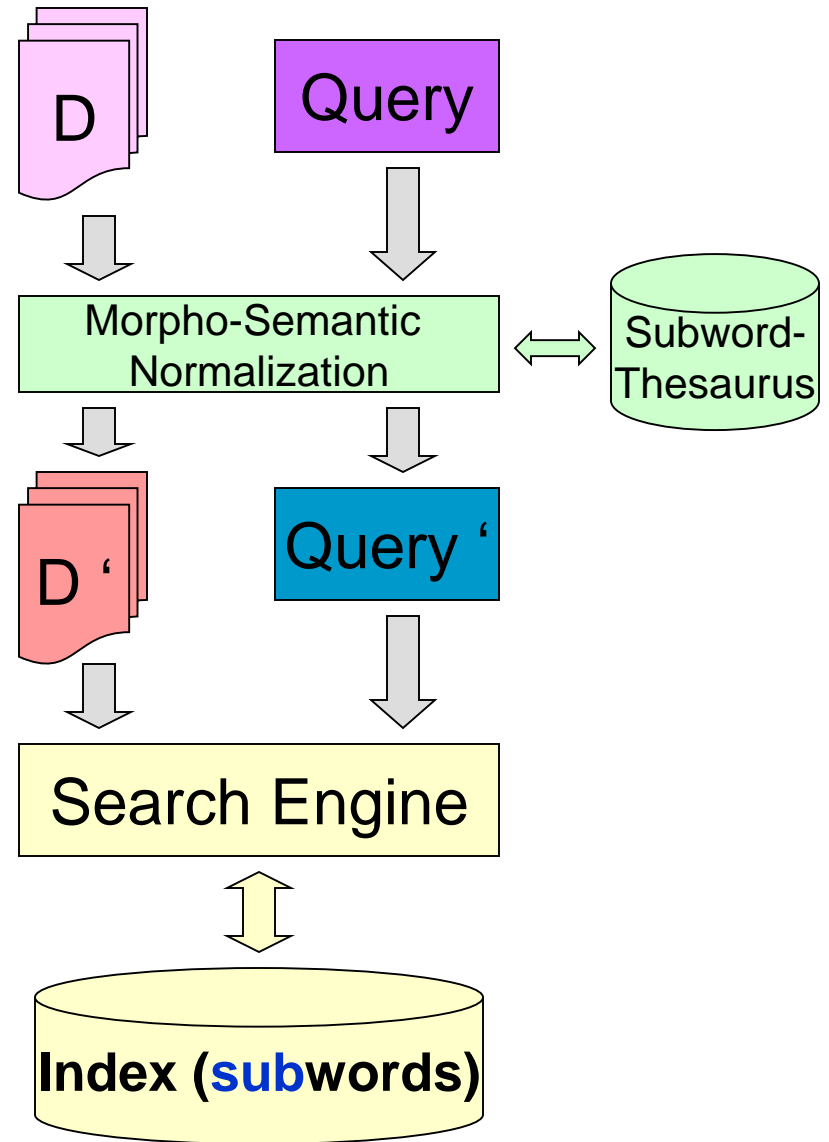
**Segmented text mapped to thesaurus lds (D')**



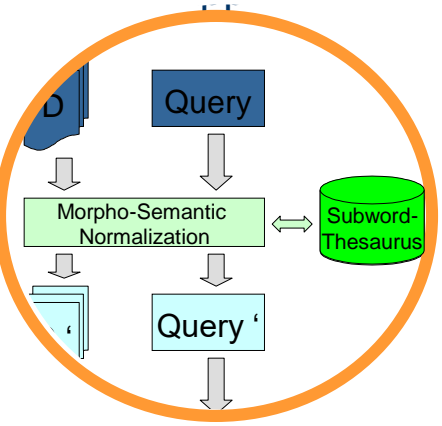
# Conventional approach



# Subword approach



# Lexical Resources

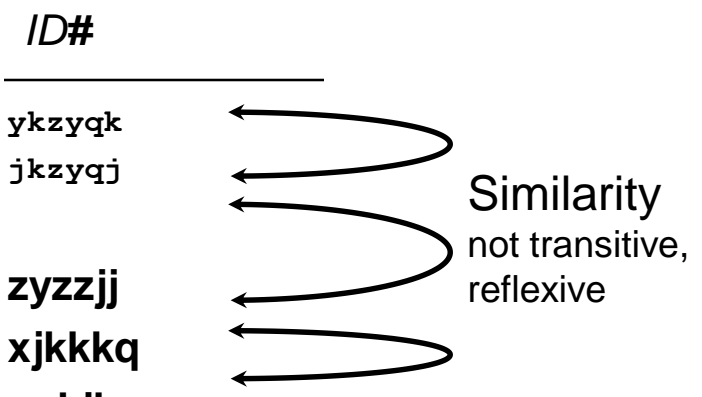


**Subword Lexicon:**  
list of subwords with attributes (type, language, etc.)

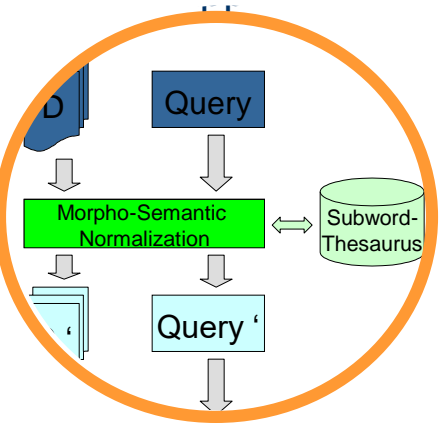
{gastr}	ykzyqk
{stomach}	jkzyqj
{magen}	
{ventric}	zyzzjj
{chamber}	xjkkkq
{hepat}	qxkjkq
{hepar}	
{liver}	
{kidney}	kkyxkj
{ren}	
{nier}	

Equivalence  
transitive  
and reflexive

**Subword Thesaurus:**  
groups **equivalent** subwords,  
links similar groups



# Algorithmic Resources



- Morphosyntactic parser based on a word model described as a finite-state automaton
- Heuristic rules for disambiguation of parses

