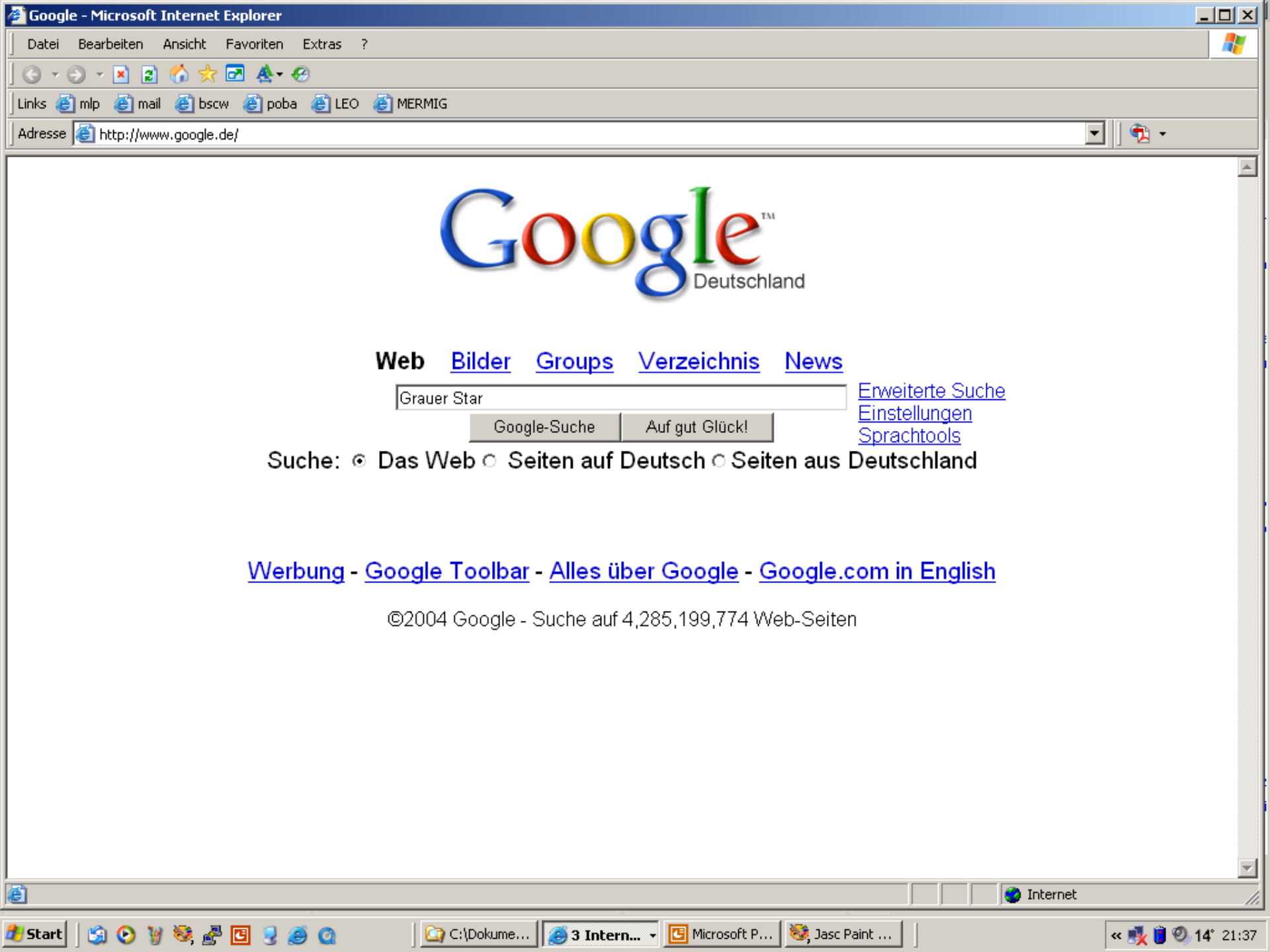
Verbesserung der Recherche in medizinischen Textkollektionen durch Wortstamm-basierte Indexierung

Stefan Schulz

Abteilung Medizinische Informatik,
Universitätsklinikum Freiburg

UNIVERSITÄTS FREIBURG KLINIKUM


Google - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Links mlp mail bscw poba LEO MERMIG
Adresse http://www.google.de/
Google™ Deutschland
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
Internet
Start
C:\Dokume...
3 Intern...
Microsoft P...
Jasc Paint ...
14° 21:37

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 - 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 - 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 - 13k

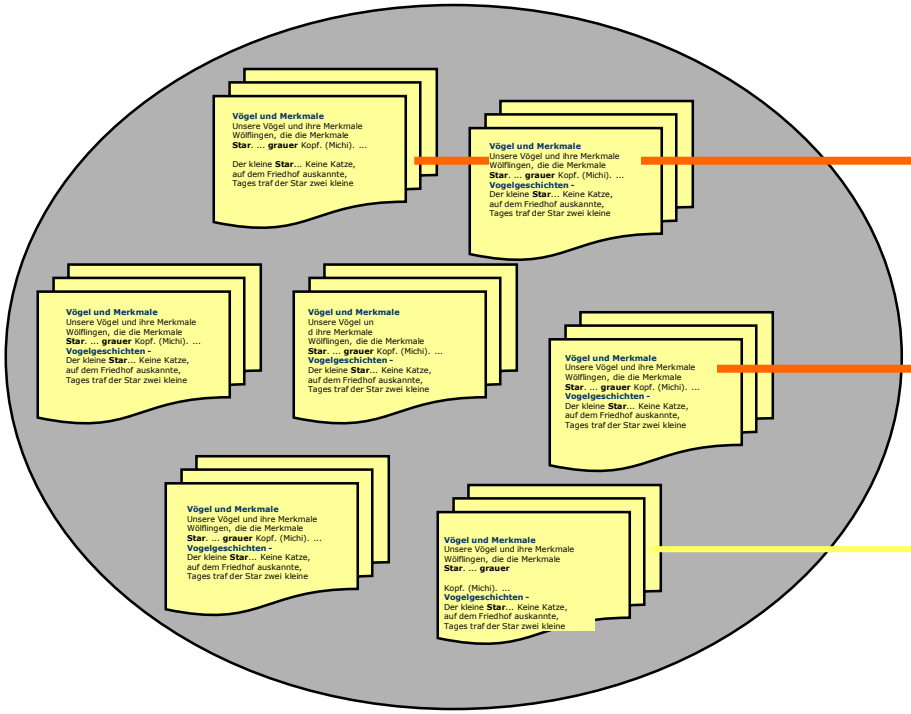
Vogelgeschichten -
Der kleine **Star**... Keine Katze, kein Hund, kein älterer **grauer** Mann ... auf dem Friedhof auskannte, bat die anderen Vögel und auch ... Eines Tages traf der Star zwei kleine Eichhörnchen ...
www.tiergeschichten.de/voegel/derkleinestar.htm - 21k -

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 - 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 -

Textretrieval



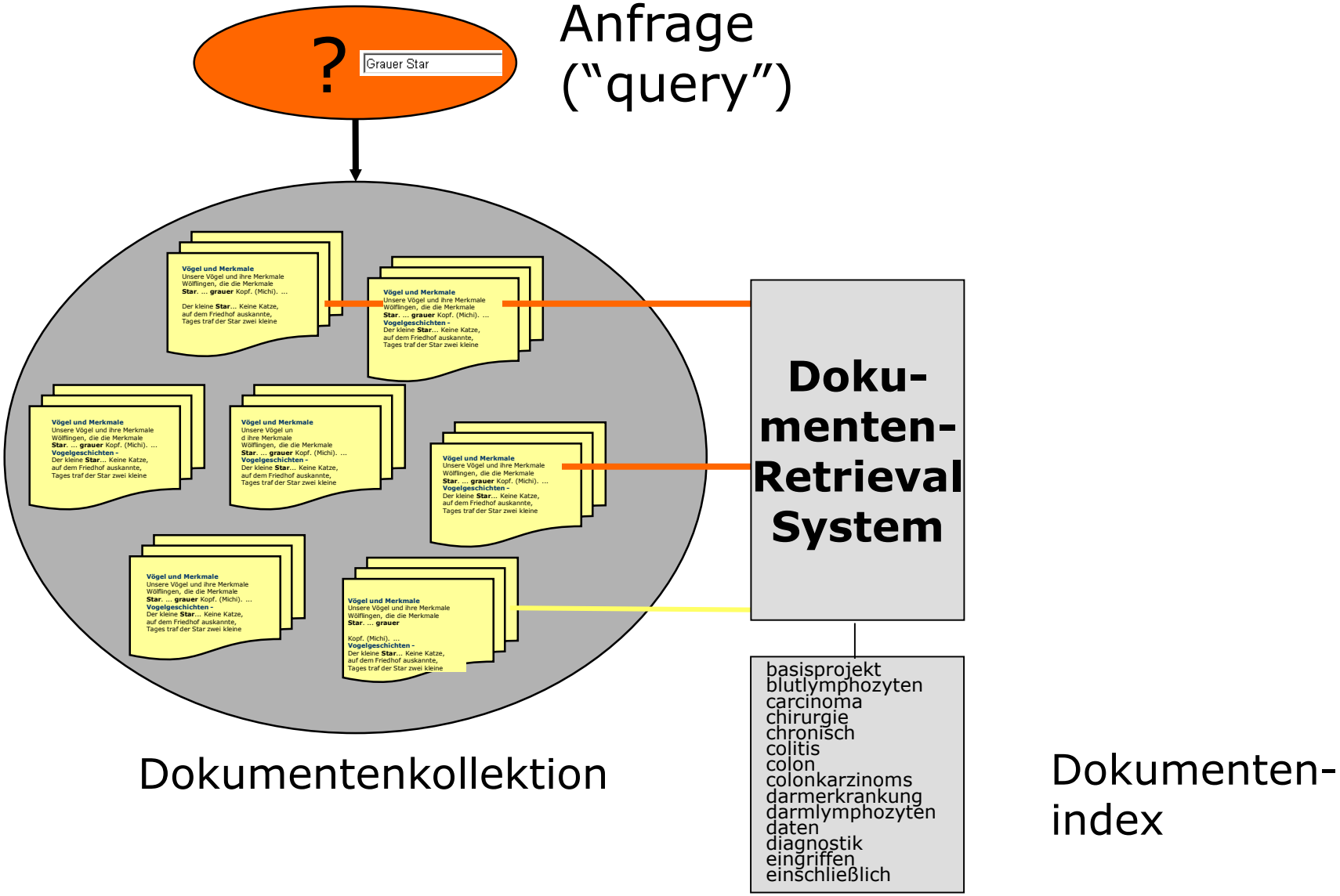
Dokumentenkollektion

Doku-menten-Retrieval System

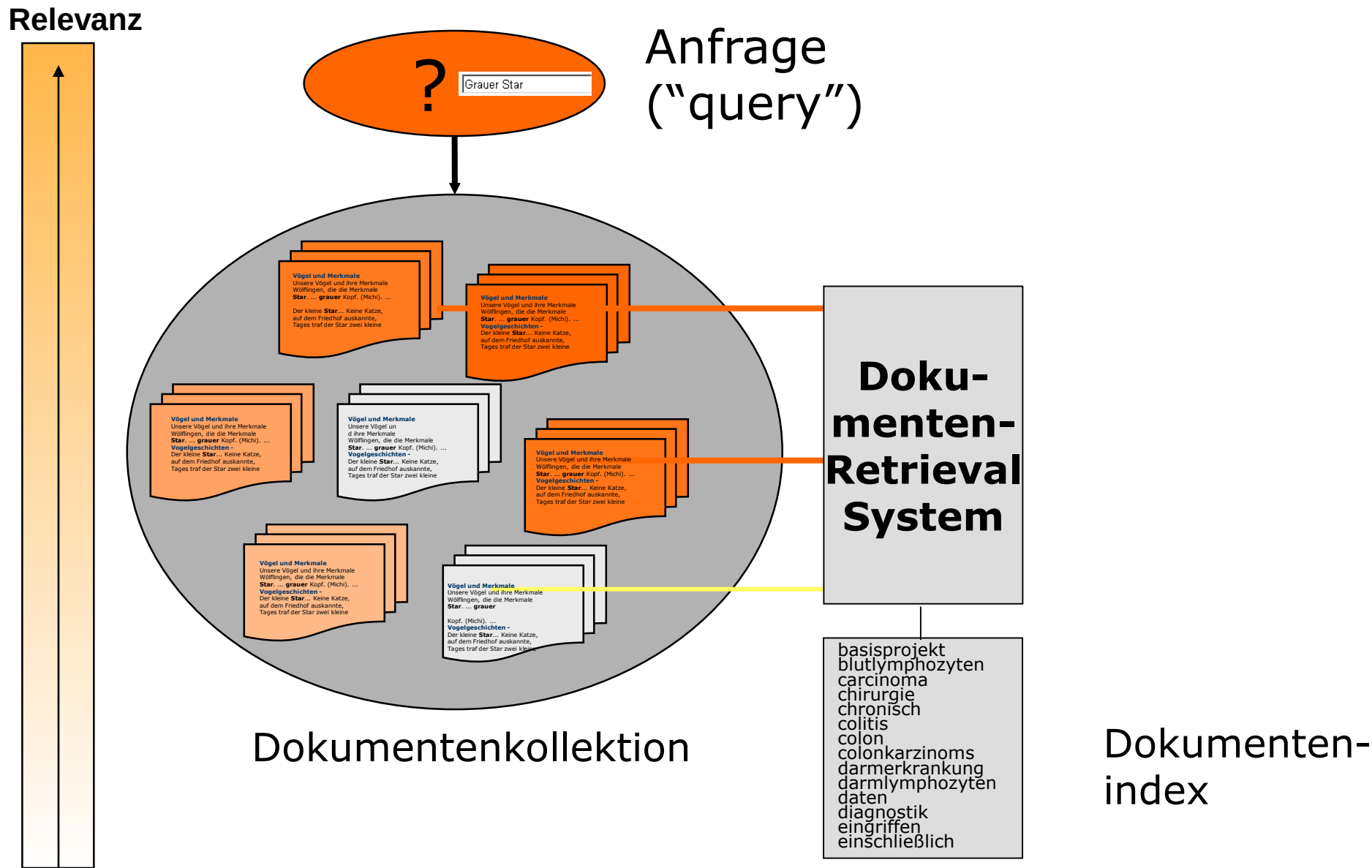
basisprojekt
blutlymphozyten
carcinoma
chirurgie
chronisch
colitis
colon
colonkarzinoms
darmerkrankung
darmlymphozyten
daten
diagnostik
eingriffen
einschließlich

Dokumenten-index

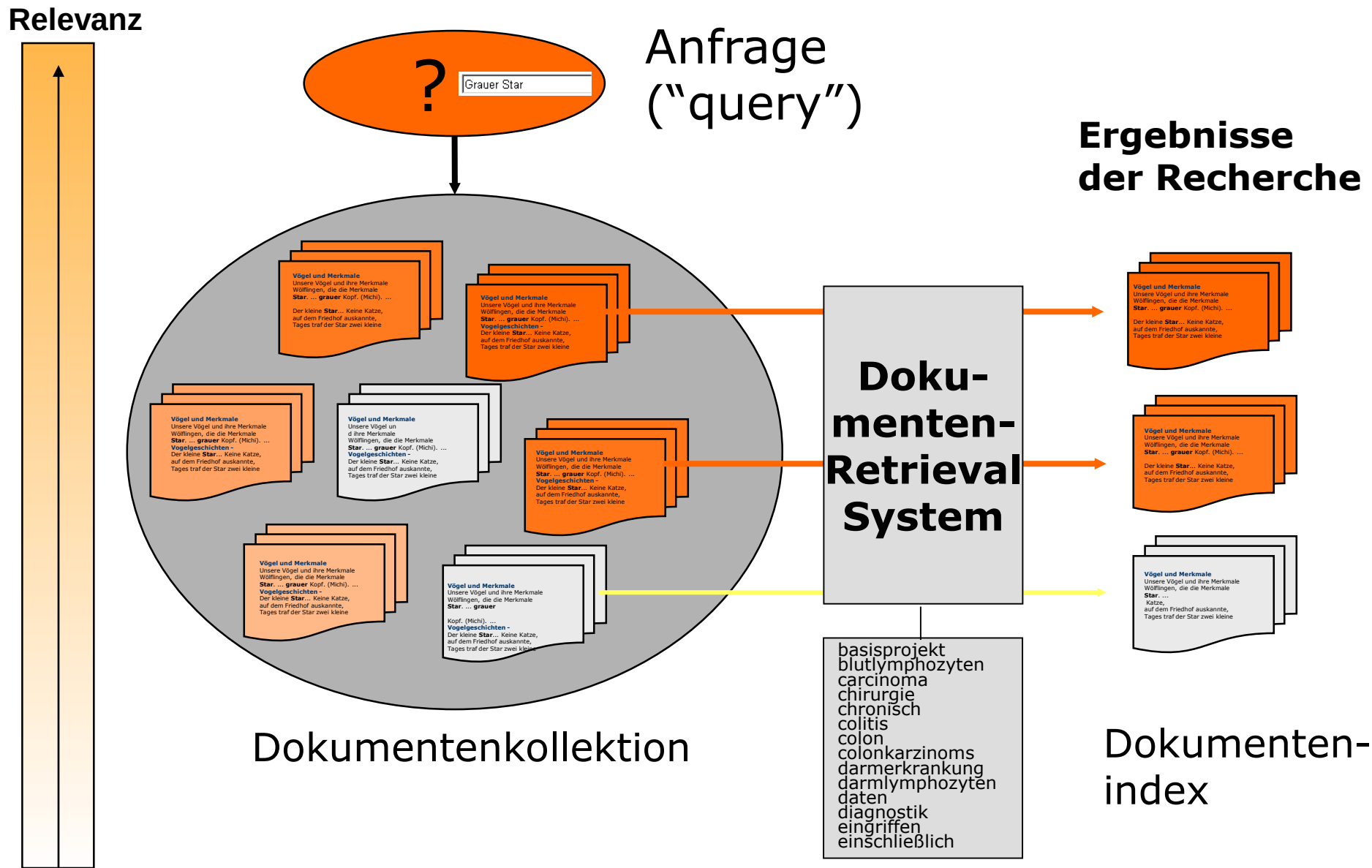
Textretrieval



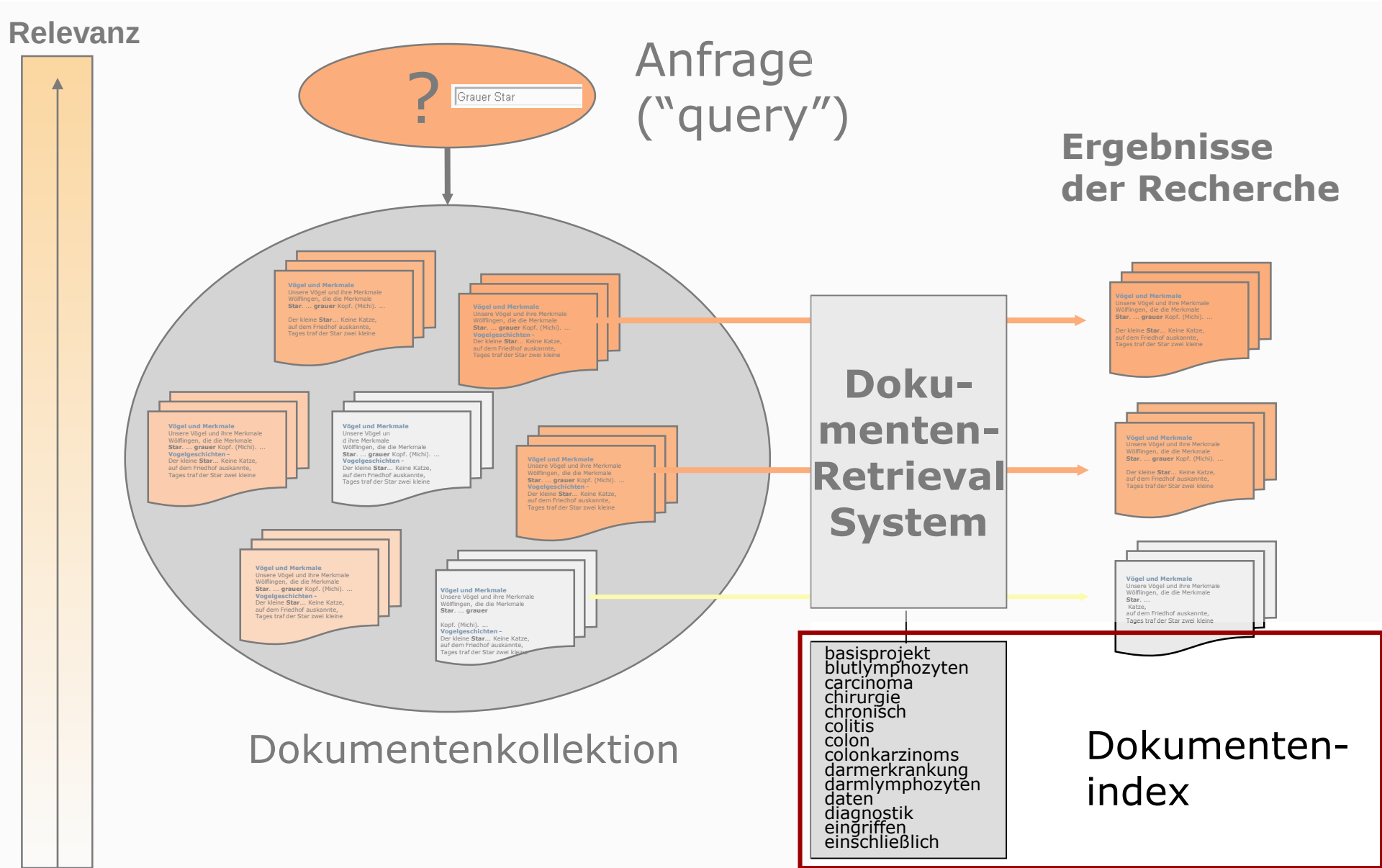
Textretrieval



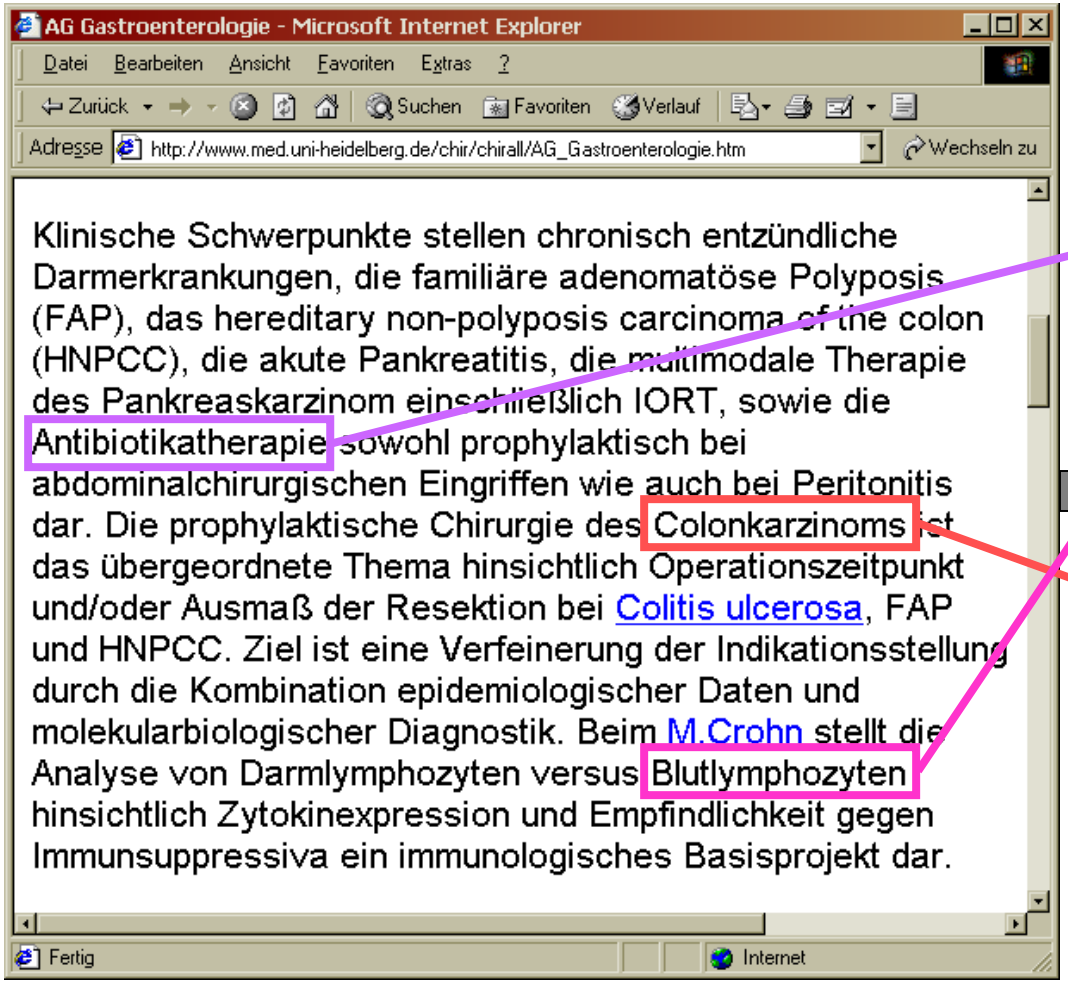
Textretrieval



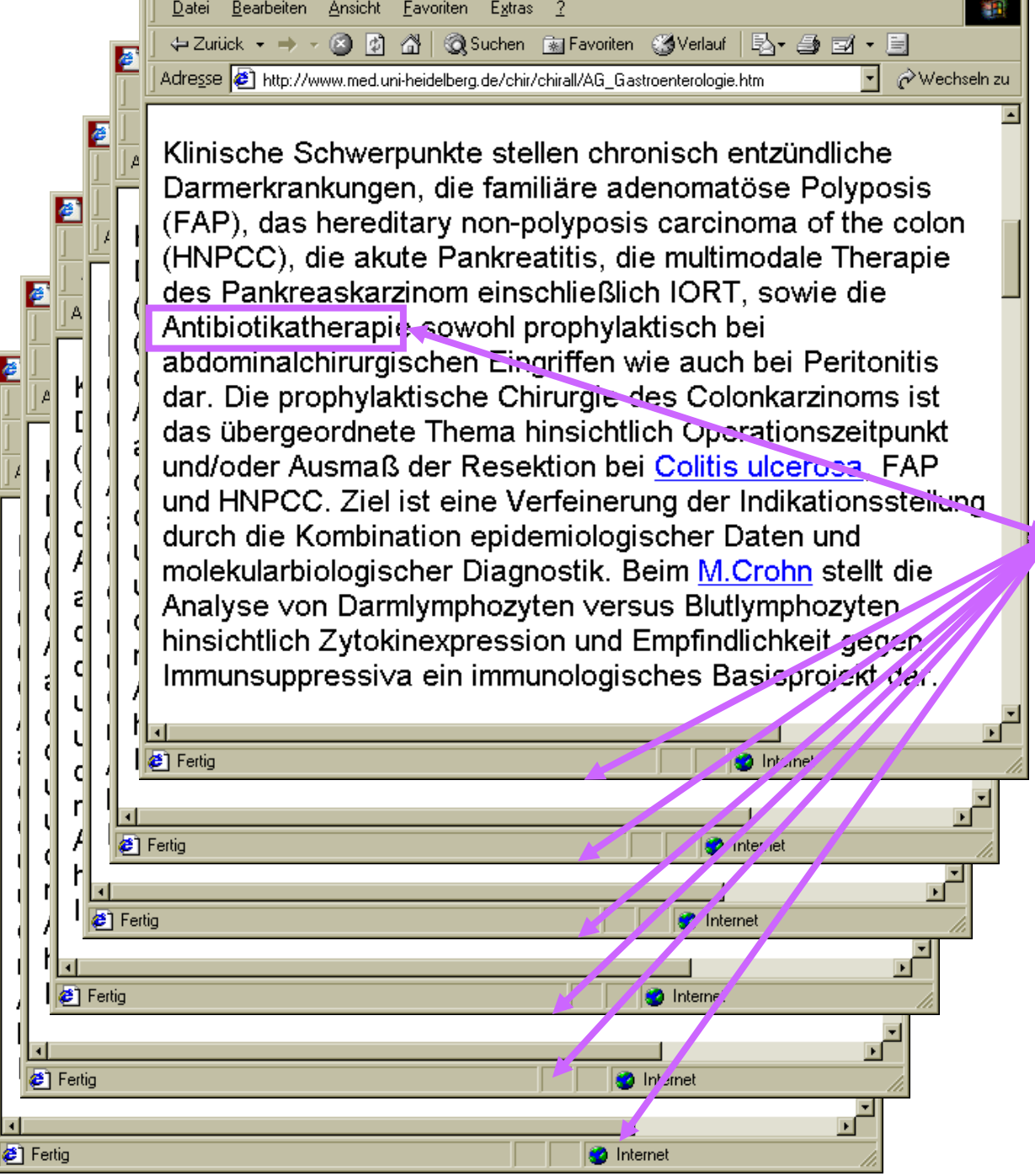
Textretrieval



Automatische Indexierung: Wortindex




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



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

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*

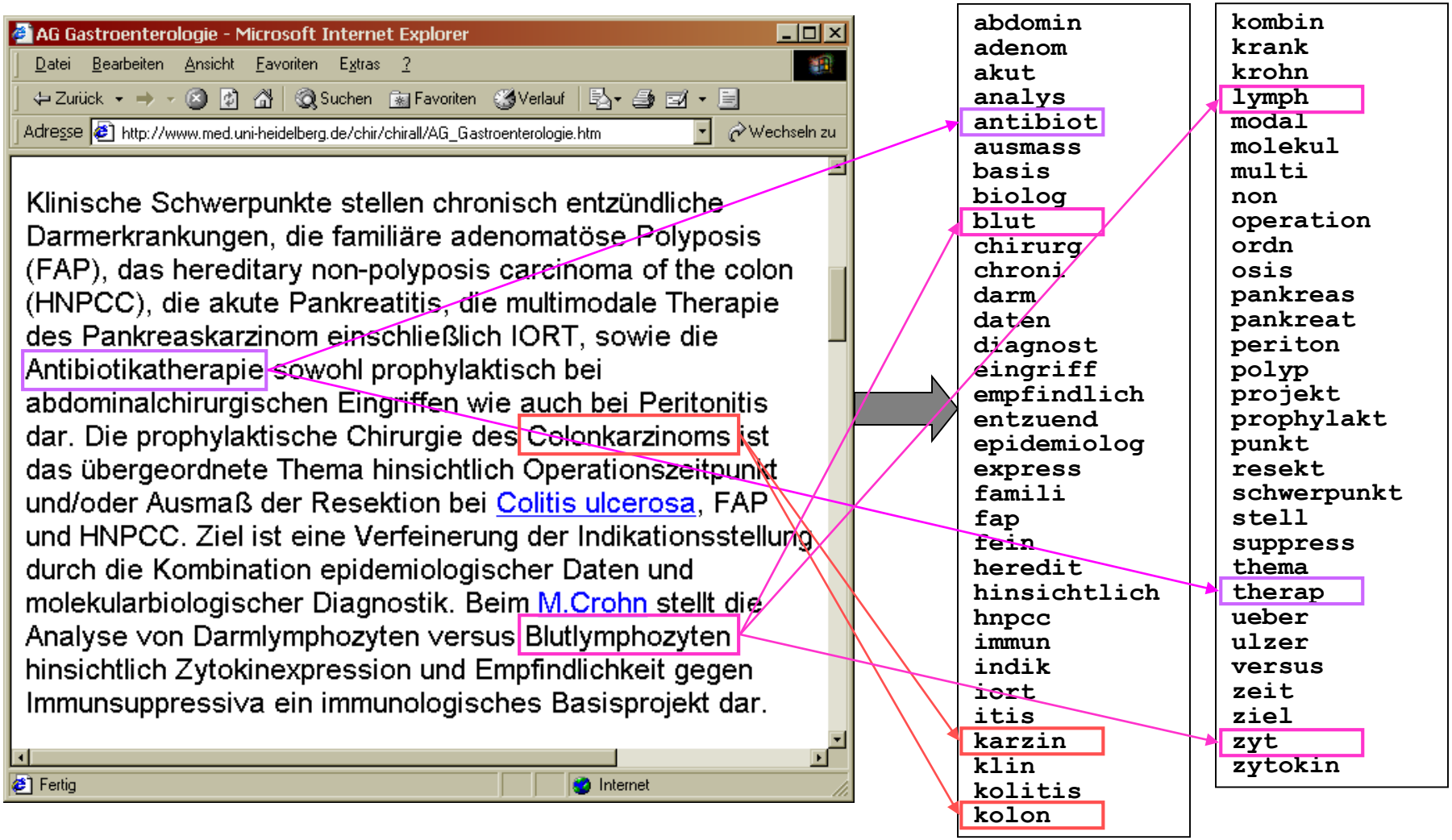
Lösungsansatz:
Subwort-Index statt Wort-Index

- 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}
  - yxwqzv = {leber, hepat, liver}

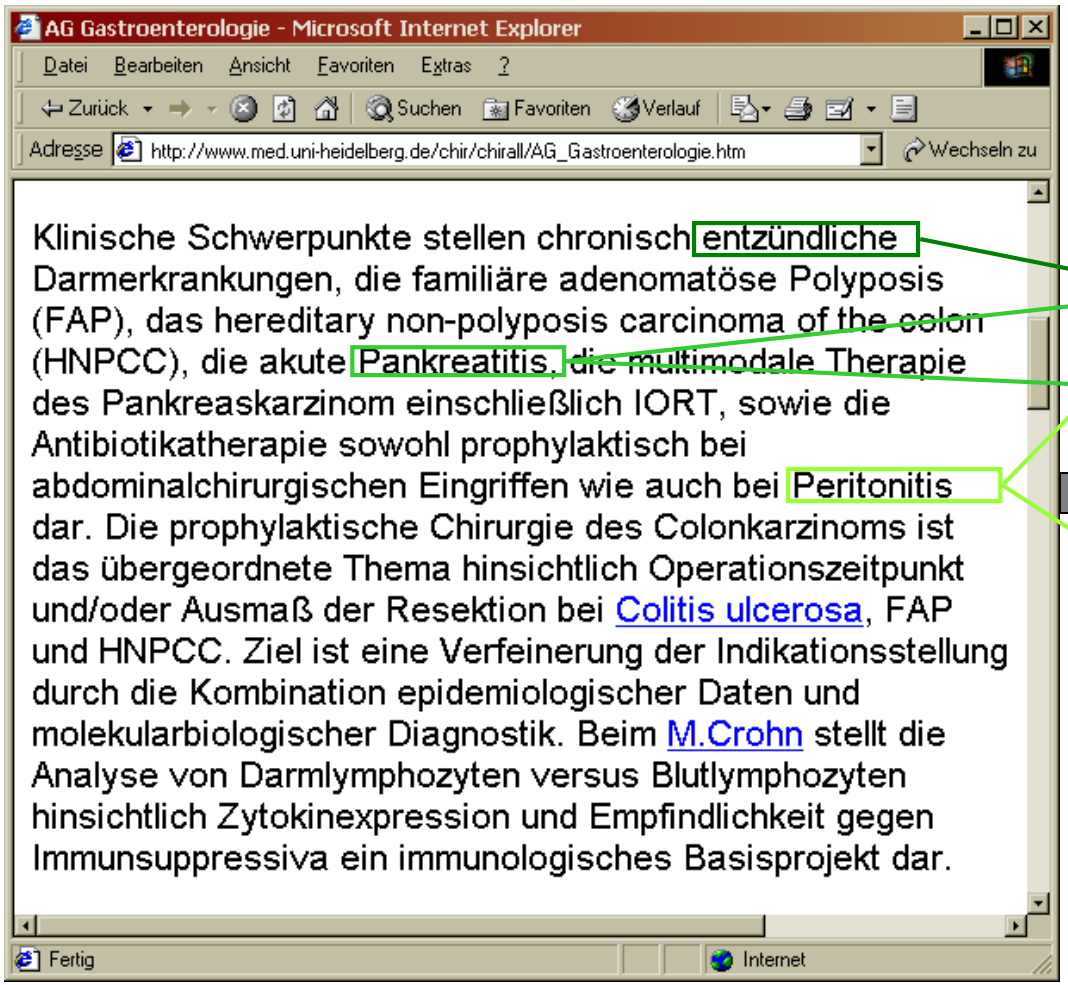
Ressourcen

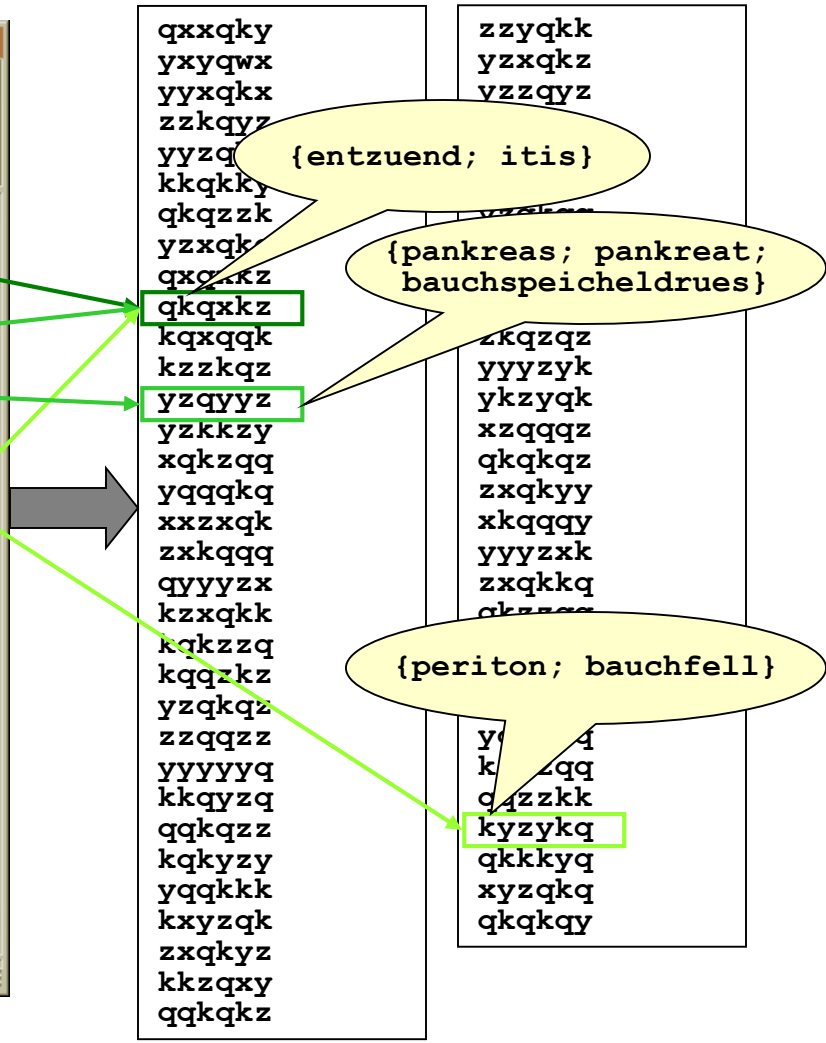
- Subwort-Lexikon:
  - Organisiert und klassifiziert medizinspezifische Subwörter und Affixe in mehreren Sprachen (derzeit Deutsch, Englisch, Portugiesisch, ca. 25.000 Einträge), Spanisch, Französisch, Schwedisch im Aufbau
- Subwort-Thesaurus:
  - Gruppiert synonyme Lexikoneinträge
- Morphosyntaktischer 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:

$$precision = \frac{n_{gefundene+relevanteDokumente}}{n_{gefundeneDokumente}}$$

$$recall = \frac{n_{gefundene+relevanteDokumente}}{n_{relevanteDokumente}}$$

- Precision/Recall-Diagramme bei geranktem Output

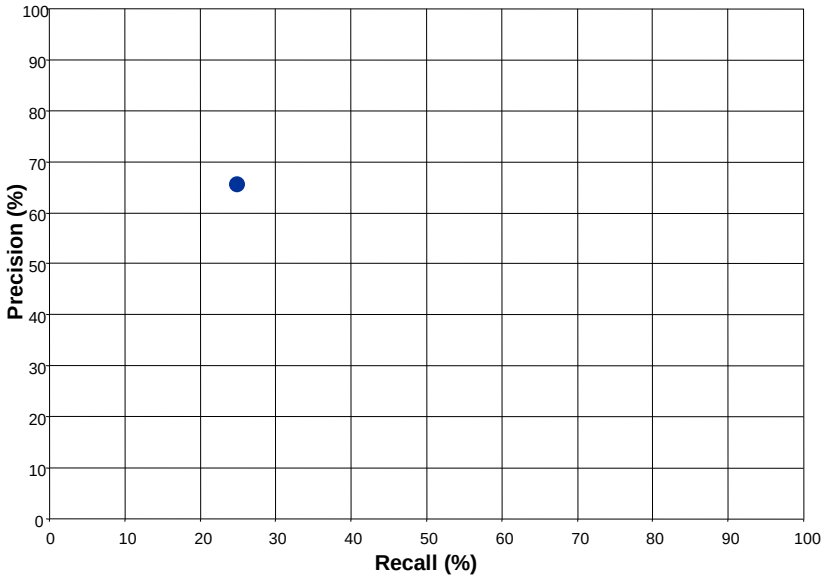
Beispiel: 25 Dokumente, 8 relevant

Anfrage X

$precision = 67\%$
$recall = 25\%$

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



Anfrage X

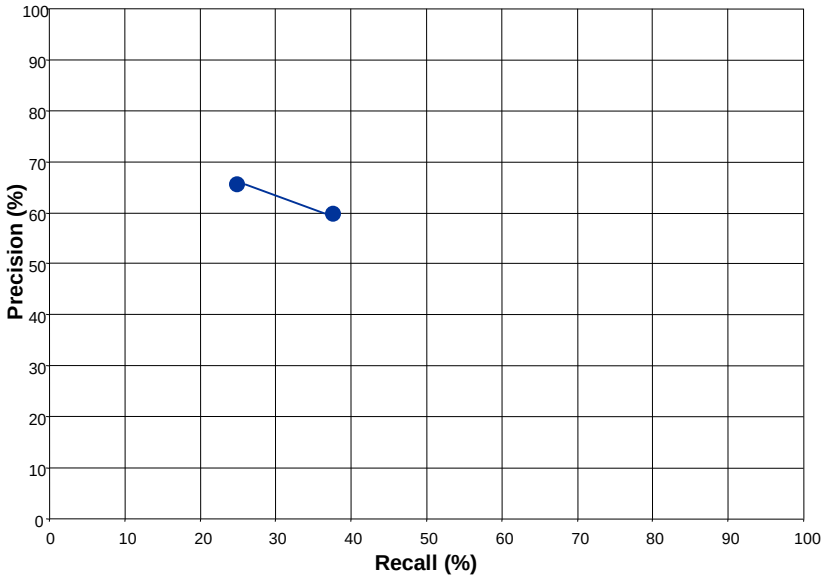
precision =
recall =

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

- Precision/Recall-Diagramme bei geranktem Output

Beispiel: 25 Dokumente, 8 relevant

Textretrievalsysteme: Evaluationsmethodik



Anfrage X

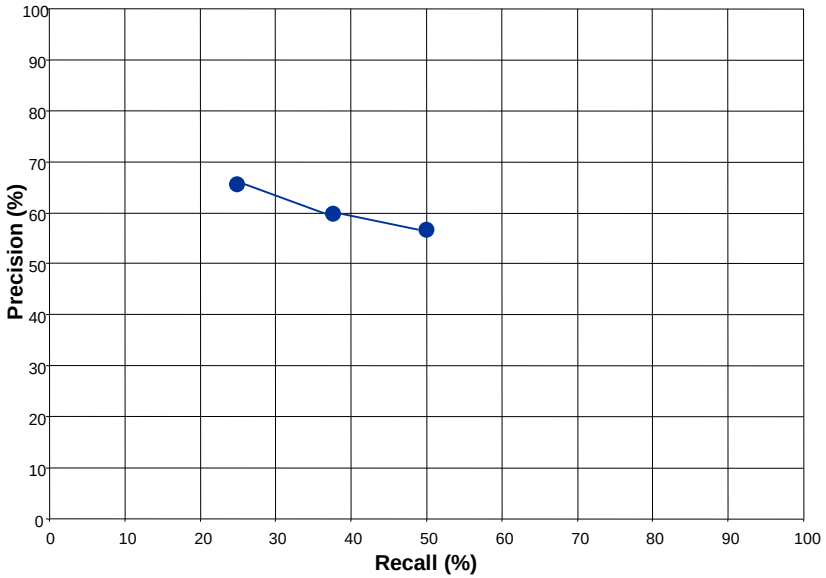
precision = 60%
recall = 38%

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

- Precision/Recall-Diagramme bei geranktem Output

Beispiel: 25 Dokumente, 8 relevant

Textretrievalsysteme: Evaluationsmethodik



Anfrage X

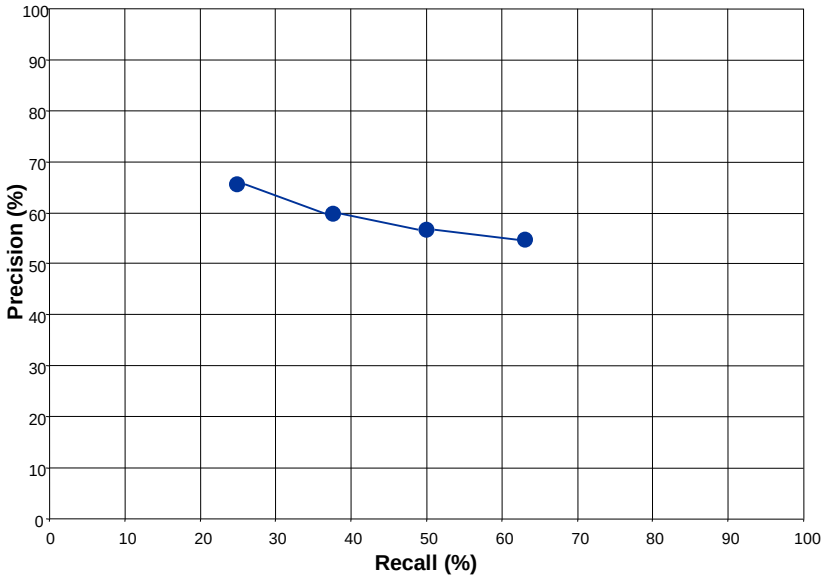
precision = *57%*
recall = *50%*

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

- Precision/Recall-Diagramme bei geranktem Output

Beispiel: 25 Dokumente, 8 relevant

Textretrievalsysteme: Evaluationsmethodik



- Precision/Recall-Diagramme bei geranktem Output

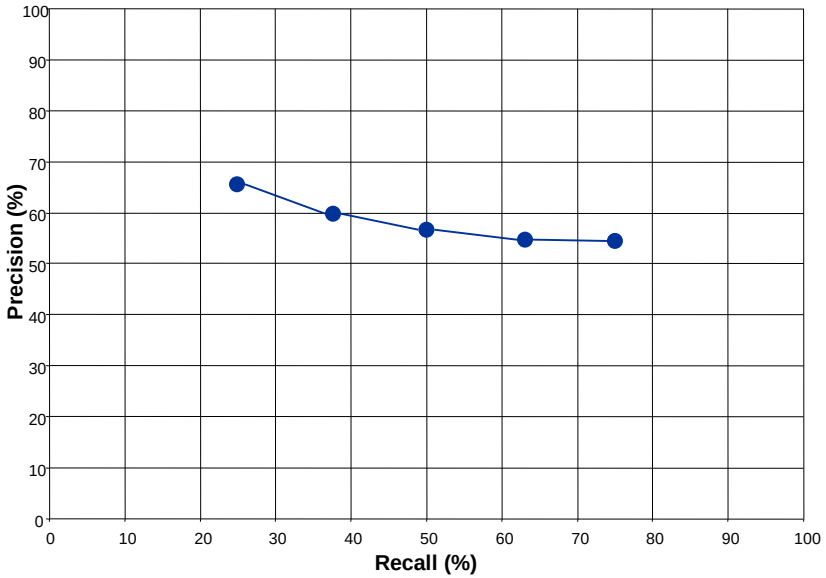
Beispiel: 25 Dokumente, 8 relevant

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%

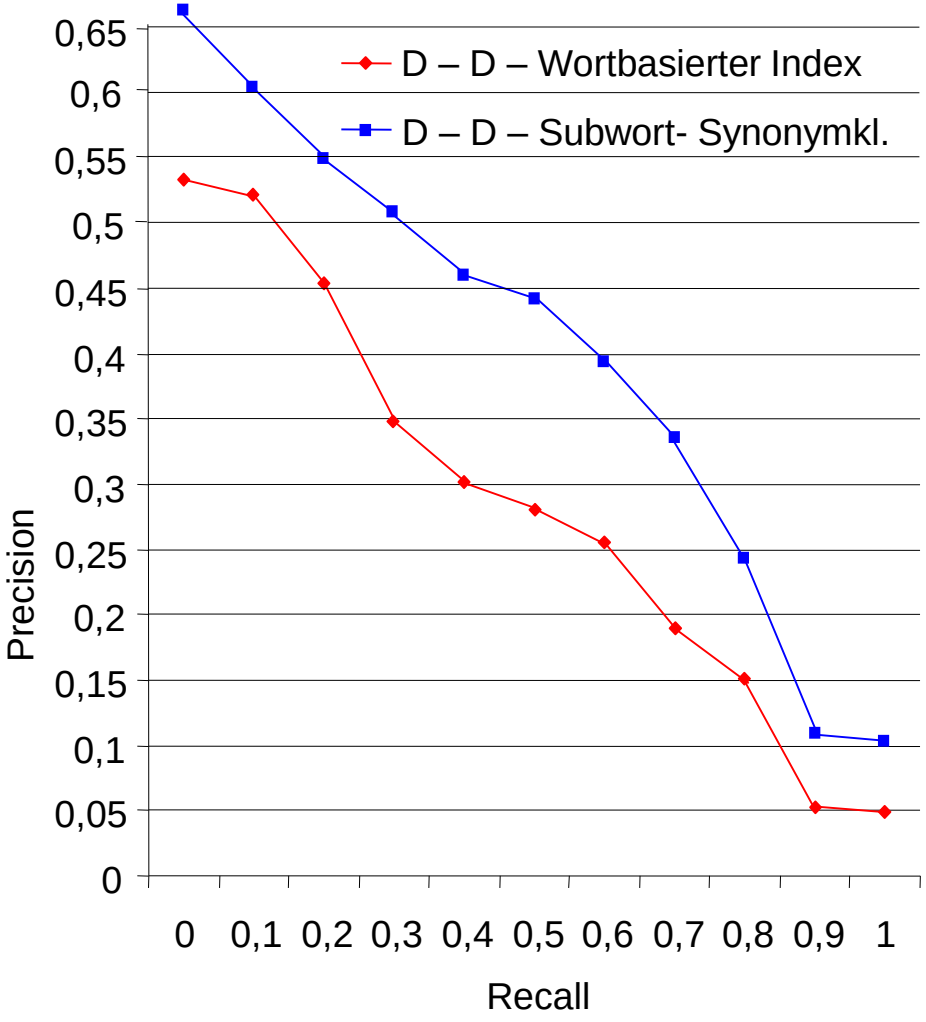
Dokument 05
Dokument 16
Dokument 21
Dokument 22
Dokument 02
Dokument 25
Dokument 20
Dokument 10
Dokument 07
Dokument 18
Dokument 04
Dokument 12
Dokument 11
Dokument 24
Dokument 15
Dokument 09
Dokument 17
Dokument 08
Dokument 19
Dokument 13
Dokument 03
Dokument 14
Dokument 23
Dokument 01
Dokument 06

Evaluationsszenarien

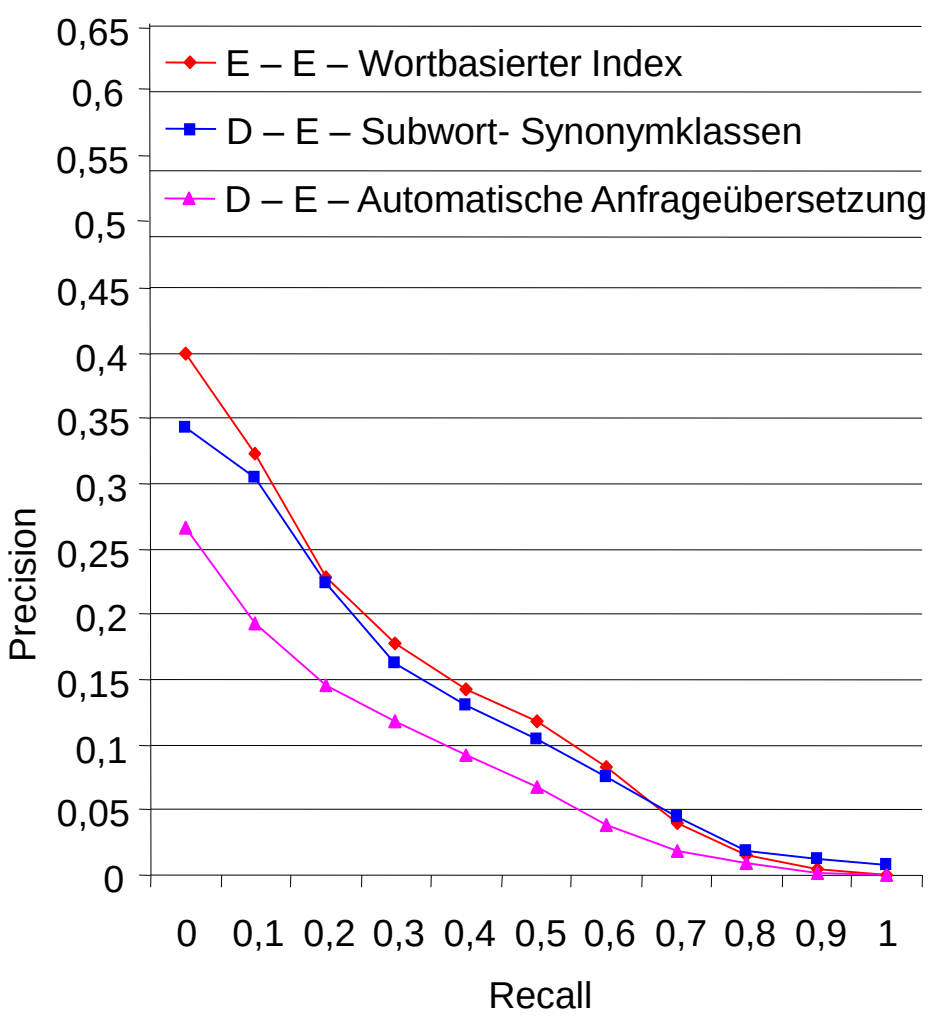
	Szenario 1	***Szenario 2***
Sprachen	D: Deutsch Q: Deutsch	D: Englisch Q: Deutsch, Englisch
Dokumente	MSD-Manual ($\|D\| = 5.500$)	MEDLINE-Abstracts ($\|D\| = 233.000$)
Anfragen	$\|Q\| = 25$ (nach IMPP-Fragen durch Medizinstudenten, Uni FR)	$\|Q\| = 106$ (Oregon Health Science Univ.) Übersetzung durch Medizin-studenten ins Deutsche
Goldstandard: $D \otimes Q \rightarrow \{rel, n.rel\}$	Relevanzurteile durch Einzelbewerung 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)
- Abdeckunggsgrad 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

www.morphosaurus.de