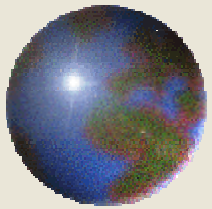
„Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt.“

Ludwig Wittgenstein
Tractatus logico-philosophicus



# *Informationssuche ohne Sprachbarrieren*

Christa Womser-Hacker

Universität Hildesheim, Informationswissenschaft

Womser@uni-hildesheim.de

www.uni-hildesheim.de/~womser

# Vortragsgliederung

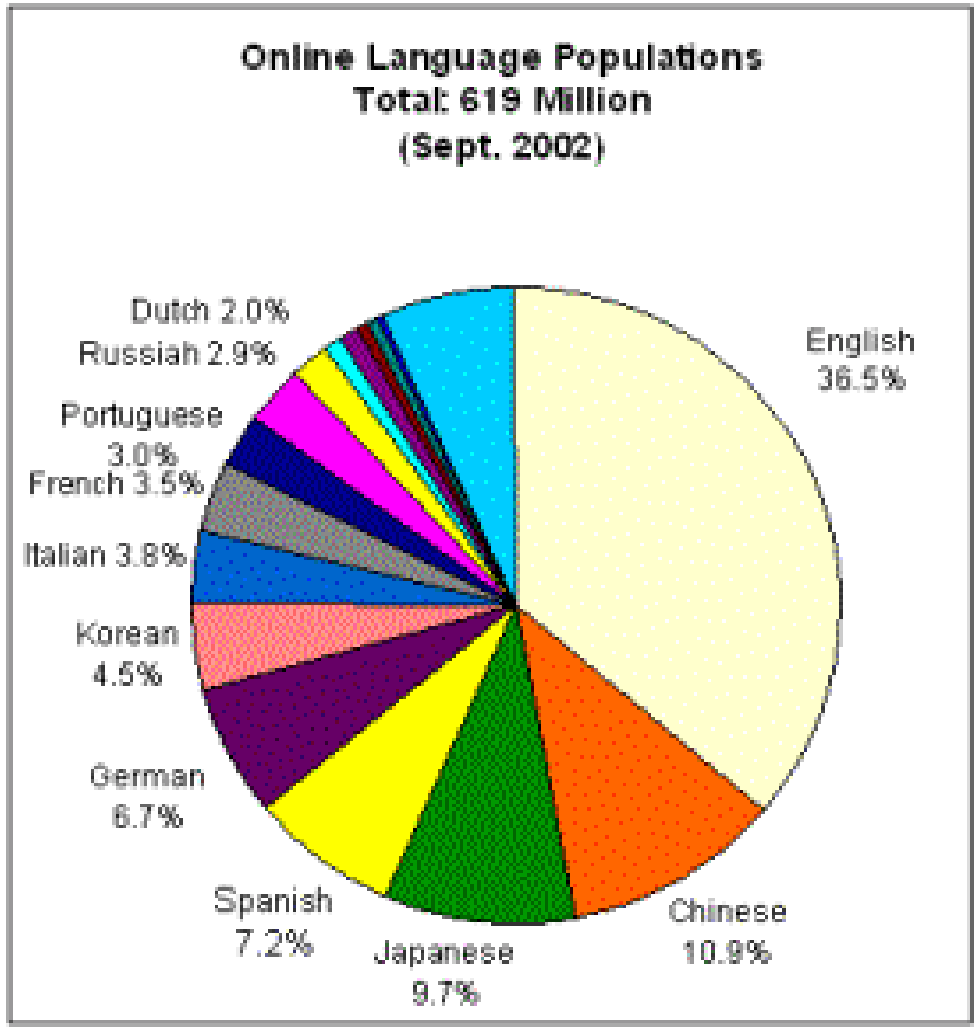
- Ausgangssituation
- Multilingualität im Information Retrieval (IR)
  - Fokus: Multi- und cross-linguale Suche
  - MIMOR-System
  - Ansätze und Verfahren im Cross-Language IR (CLIR)
  - Sprachtechnologie
  - Evaluierung im Cross-Language Evaluation Forum
- Fazit und Ausblick

# Ausgangssituation

- Anwachsen internationaler Beziehungen und internationaler Projekte (EU-Projekte, NSF-EU-Kooperation)
- Global agierende Unternehmen und Organisationen
- Electronic Commerce über Sprachgrenzen hinweg eröffnet neue Märkte
- Forderung nach professionellem Wissens-management

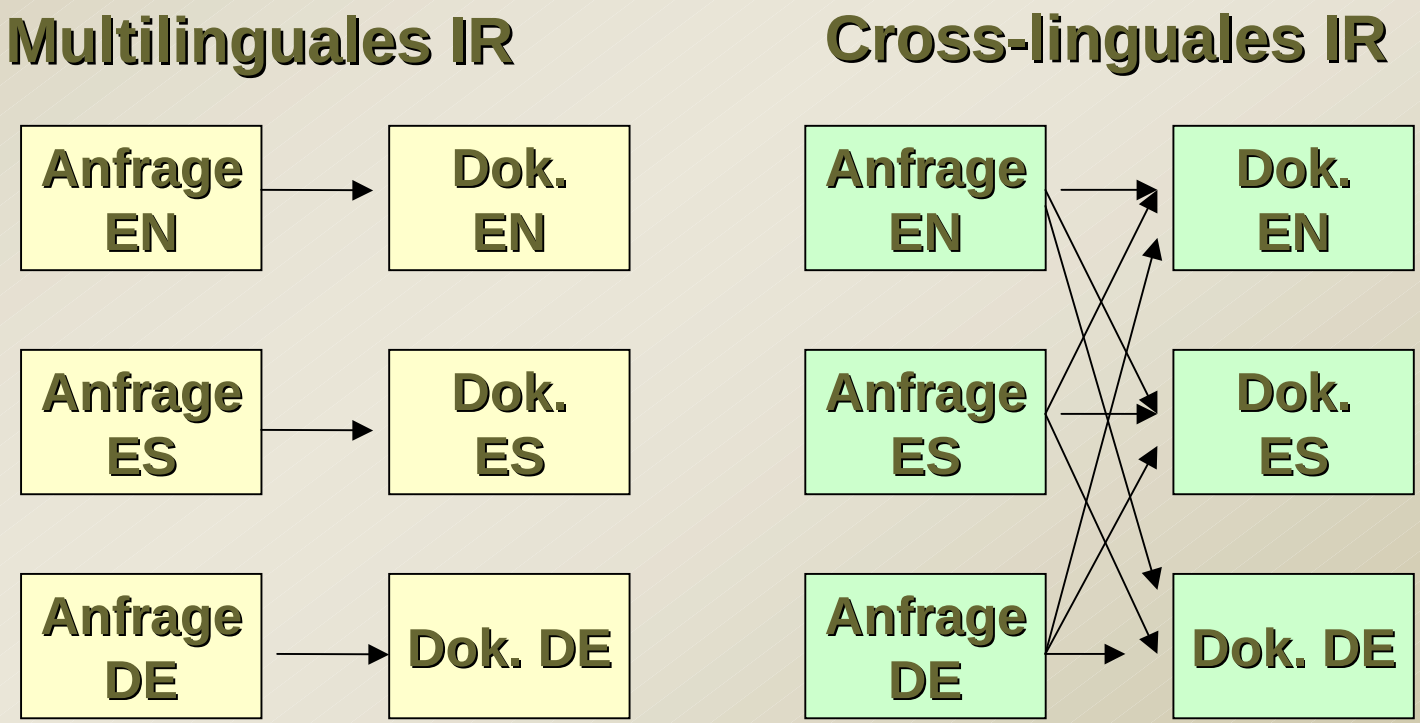
➢ Schlüsselfunktion der Mehrsprachigkeit und interkulturellen Handlungsfähigkeit

# Sprachverteilung im Internet

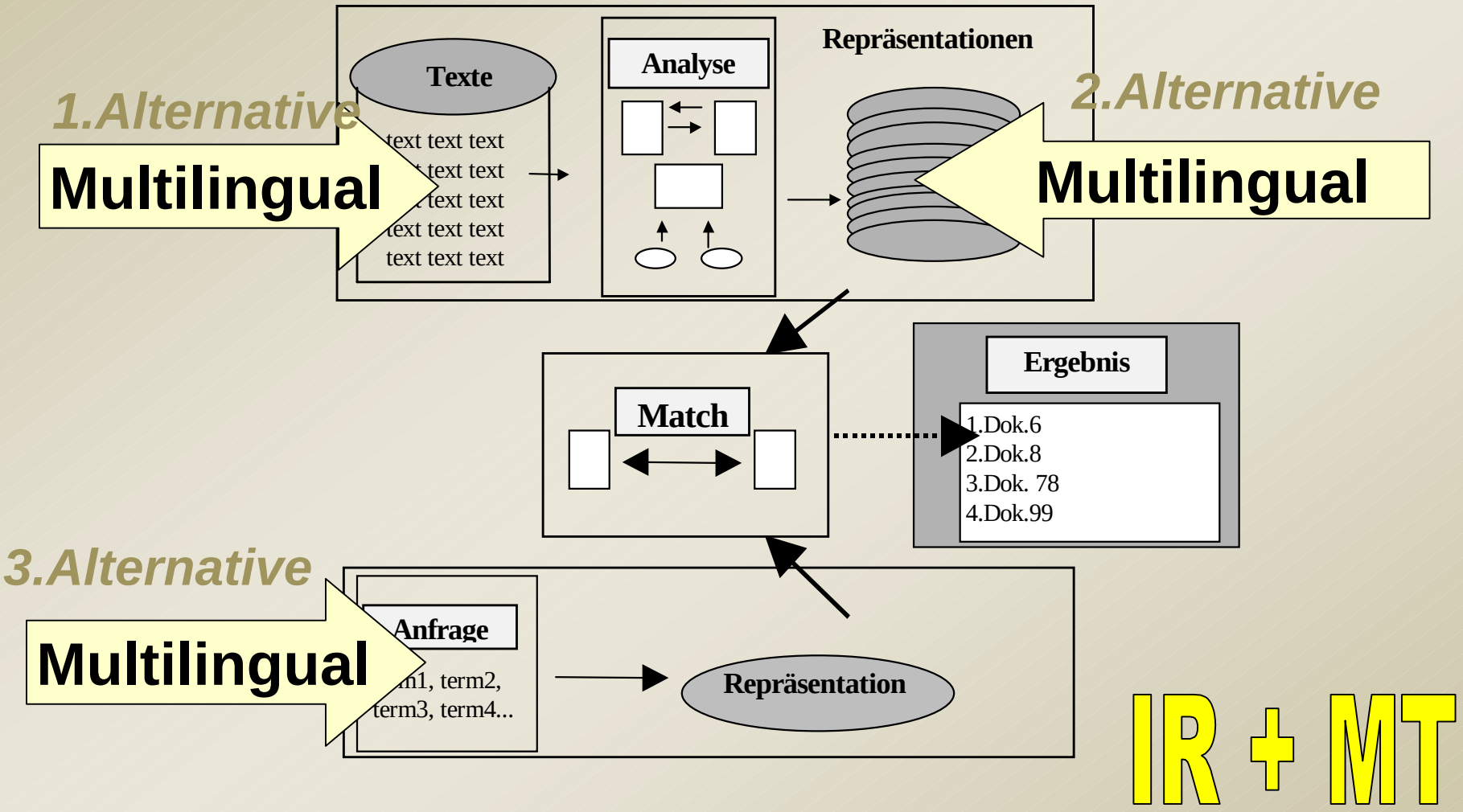



**Quelle:**
**www.hollyroodtranslations.com**

# Mehrsprachiges Information Retrieval



# Multi- und cross-linguales IR

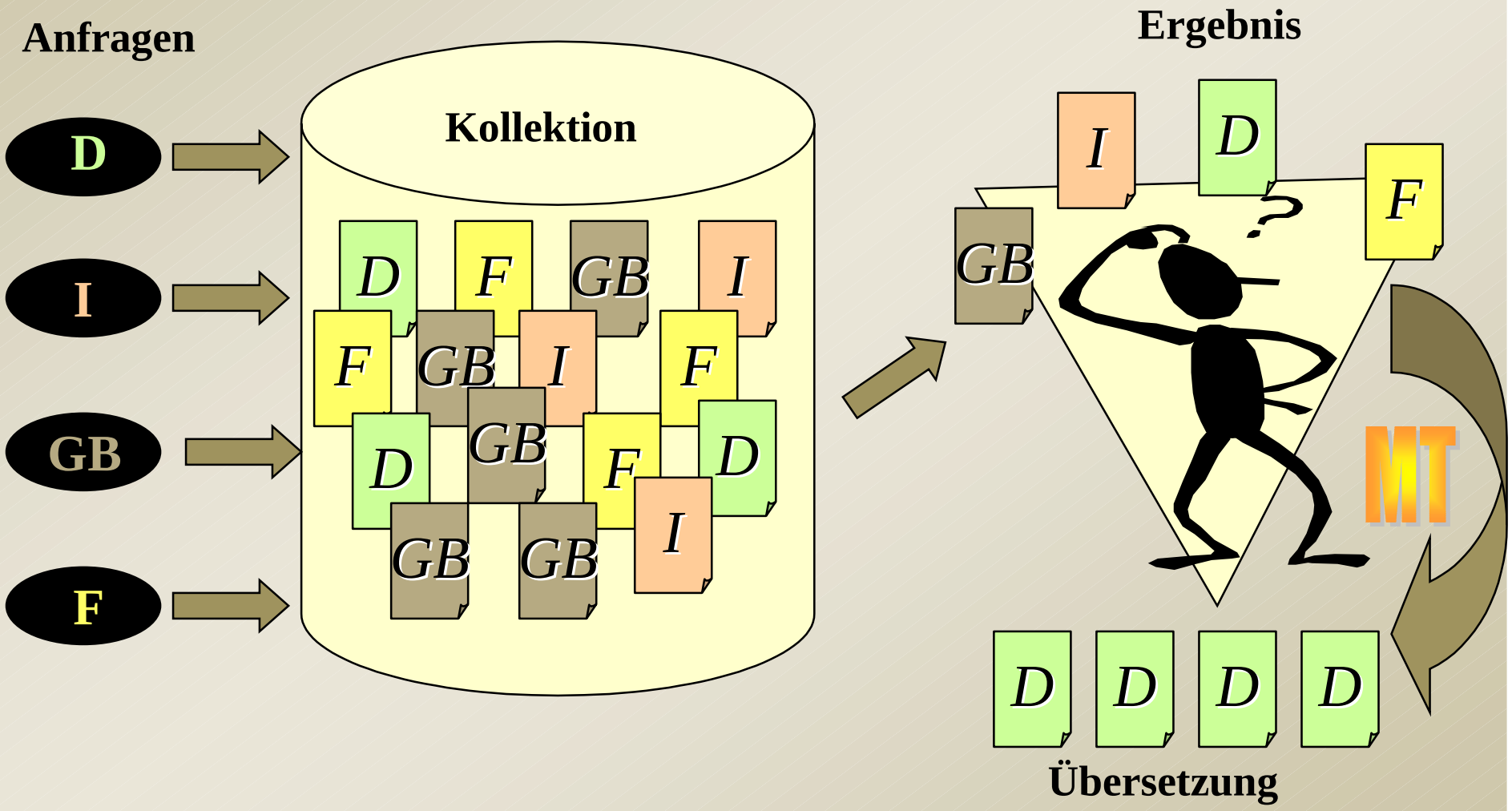



IR + MT

# Basisansätze

- Übersetzung der Dokumente
  - ressourcenintensiv
  - etwas bessere Ergebnisse [Chen / Gey 2003]
- Übersetzung der Anfrage, aber ...

# Cross-Language Information Retrieval



# *MIMOR framework*

- **„Mehrfachindexierung durch dynamische Methoden-Objekt-Relationierung“**
- Eigenentwicklung in JAVA
- Metasuche und Integration vorhandener Werkzeuge und Ressourcen
- Offenes System für Forschung und Lehre

- Anwendungsbereiche:
  - Monolinguales IR (D, E)
  - Fachsprachen
  - CLIR (D, E, F, S, I, Fi, SV, Ru)

  In Arbeit: Japanisch, Arabisch
  - Musikretrieval

# *MIMOR basics*

- Objekt-Methoden-Relationierung durch internes Bewertungsmodell
- Fusionsansatz
  - Integration verschiedener IRS bzw. Ergebnislisten
  - Integration verschiedener MT-Systeme
  - Integration verschiedener Lexika und anderer Ressourcen
- Query Translation Strategy
- Query Expansion durch Blind Relevance Feedback

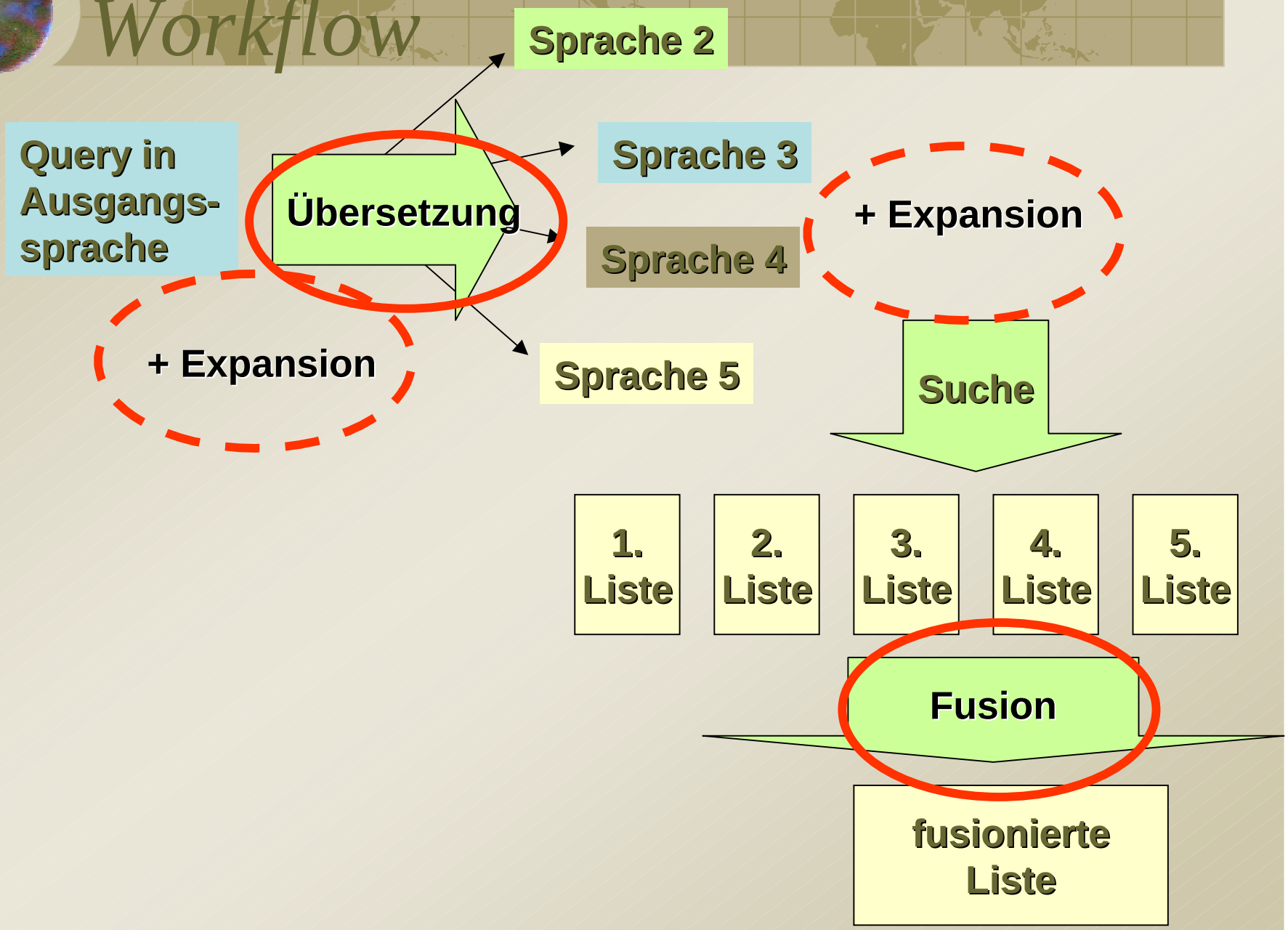
# Übersetzungsstrategie

- Entscheidung für Ausgangssprache
  - sehr oft Englisch aufgrund der vorliegenden Ressourcen
- Expansion auf der Basis der Anfrage in der Ausgangssprache
- Entscheidung für MT-Systeme (unterschiedliche Qualität pro Sprache)
- wenn kein MT-System für bestimmtes Sprachpaar verfügbar, über Pivot-Sprache (I→E→ JP)

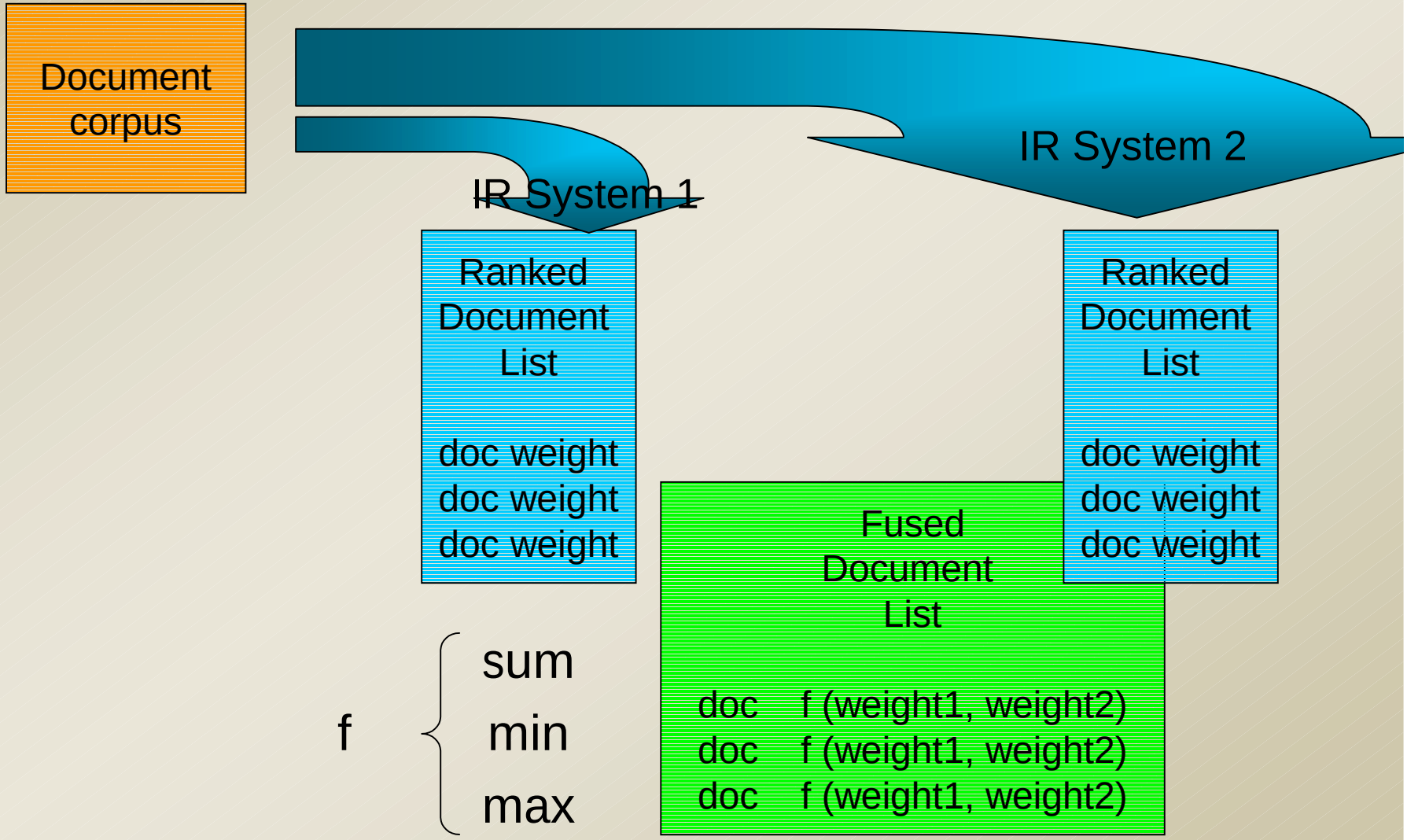
# Expansionsverfahren

- beste Qualität: Relevance Feedback Verfahren
  - aufwändig wg. Benutzerpartizipation / Interpretation
- Blind Feedback / Pseudofeedback
  - 1. Durchgang erzeugt Ranking-Liste
  - Im 2. Durchgang wird die Anfrage durch die besten Terme der besten Dokumente ergänzt (15/5)
  - Standard Rocchio

# Workflow

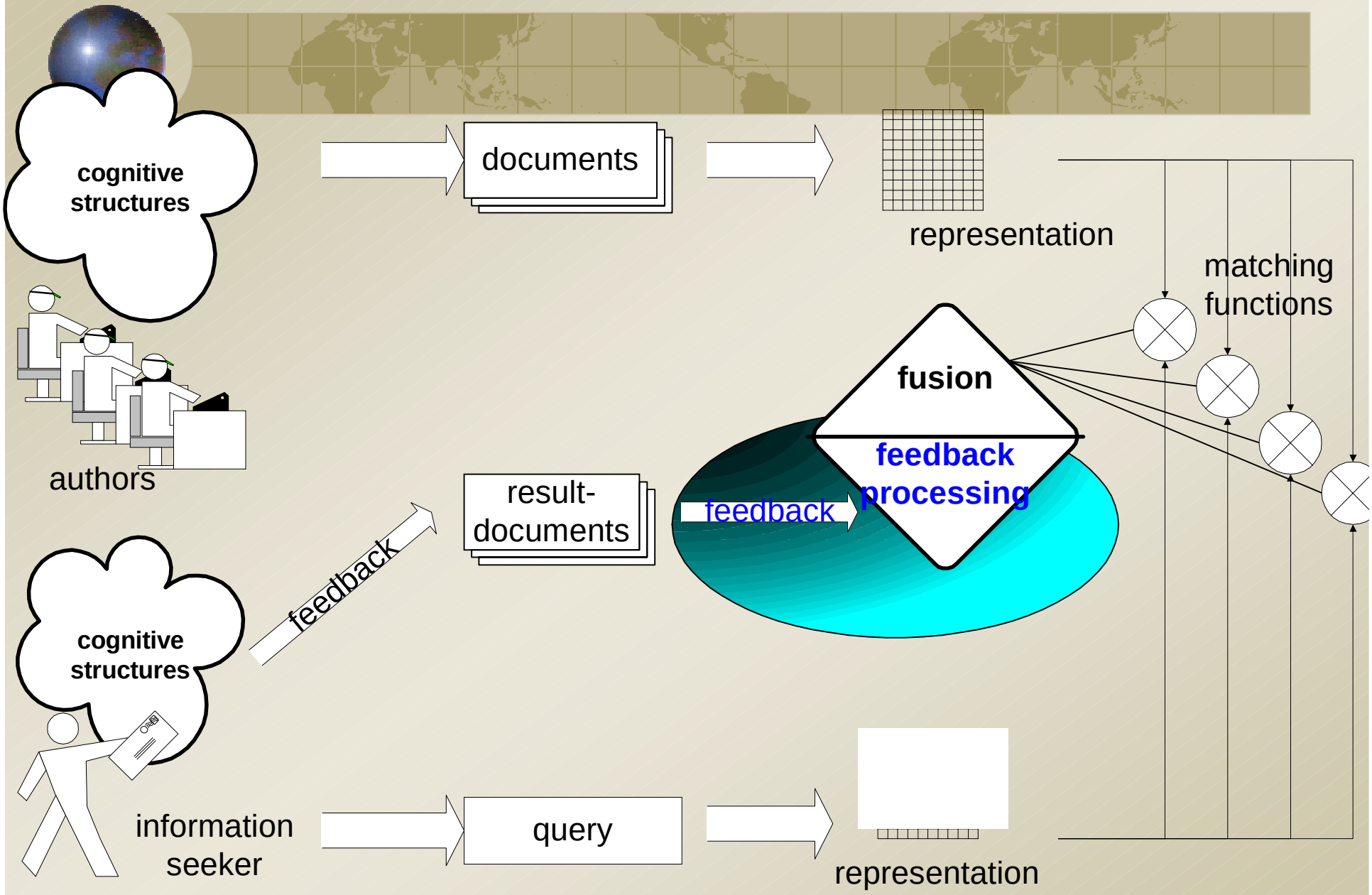

Sprache 2
Query in Ausgangs-sprache
Übersetzung
Sprache 3
Sprache 4
+ Expansion
+ Expansion
Sprache 5
Suche
1. Liste
2. Liste
3. Liste
4. Liste
5. Liste
Fusion
fusionierte Liste

# Fusion

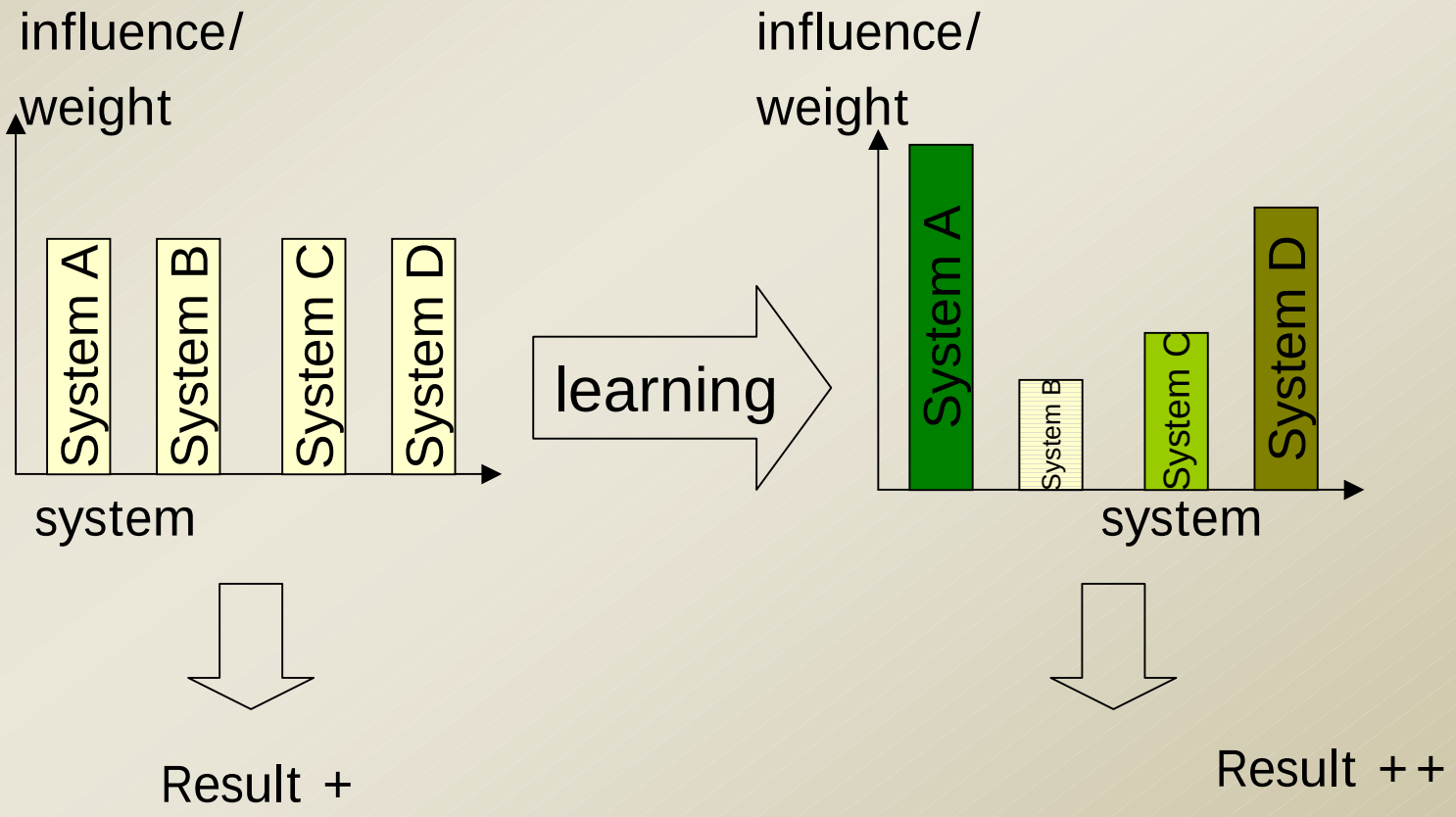

Document corpus
IR System 1
IR System 2
Ranked Document List
doc weight
doc weight
doc weight
Ranked Document List
doc weight
doc weight
doc weight
Fused Document List
doc f (weight1, weight2)
doc f (weight1, weight2)
doc f (weight1, weight2)
f
sum
min
max

# Fusionsstrategien

- round-robin
  - Reihum 1 Dokument von jeder Liste
- raw-score merging
  - Sortierung nach den berechneten RSVs
- normalized RSV
  - Sortierung nach den normalisierten RSVs
- Logistic regression
  - Neutralisierung von Ausreißern
- mean RSV
  - Sortierung nach durchschnittlichen RSVs


cognitive structures
documents
representation
matching functions
fusion
feedback processing
authors
result-documents
feedback
feedback
cognitive structures
information seeker
query
representation

# Lernprozess / Systemgewichtung



# Experimente in CLEF (1)

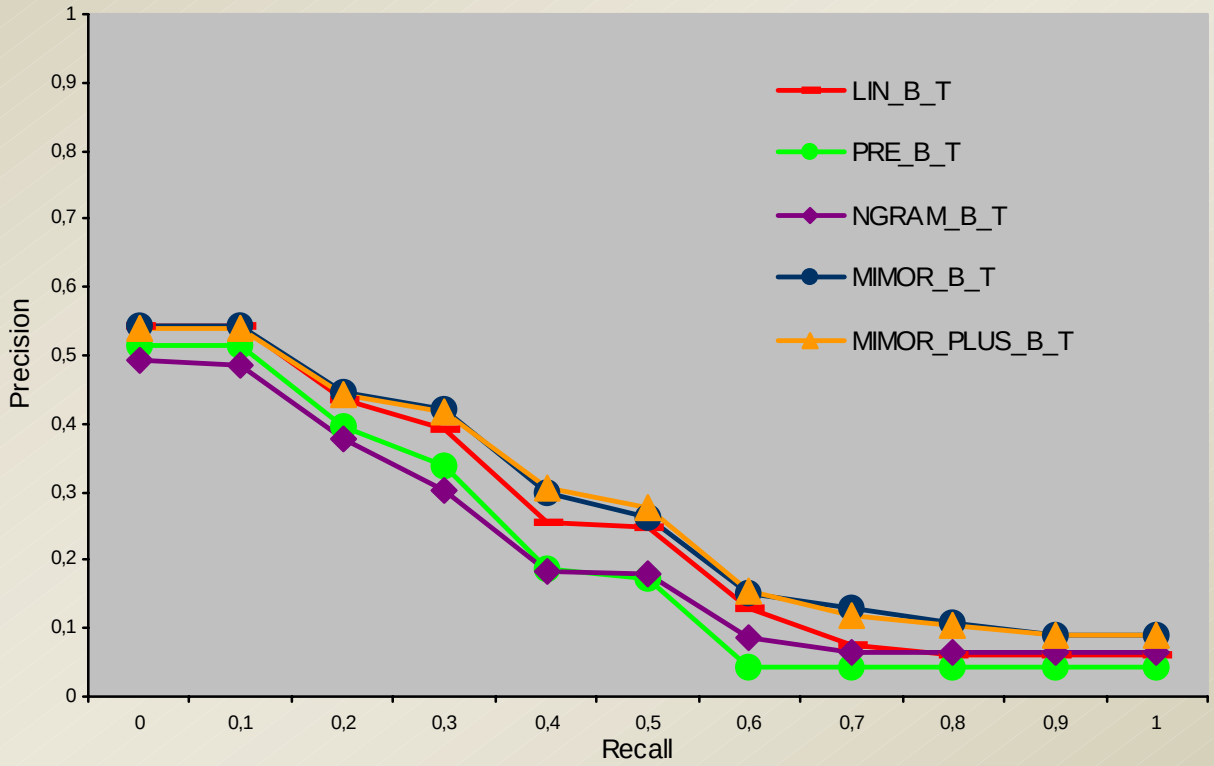



**Fig. 1 Recall-Precision-Graph for Boolean DB2 runs with** *DER SPIEGEL*

# Experimente in CLEF (2)

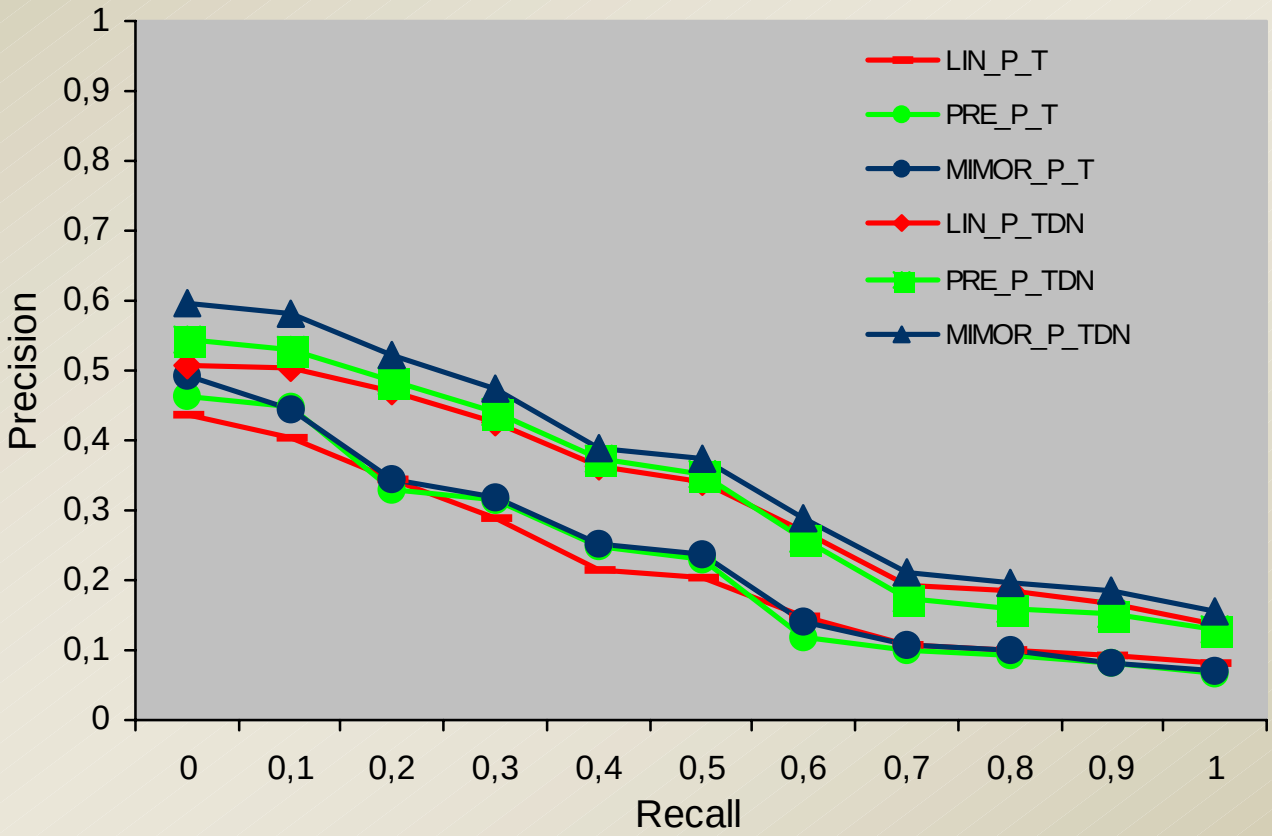
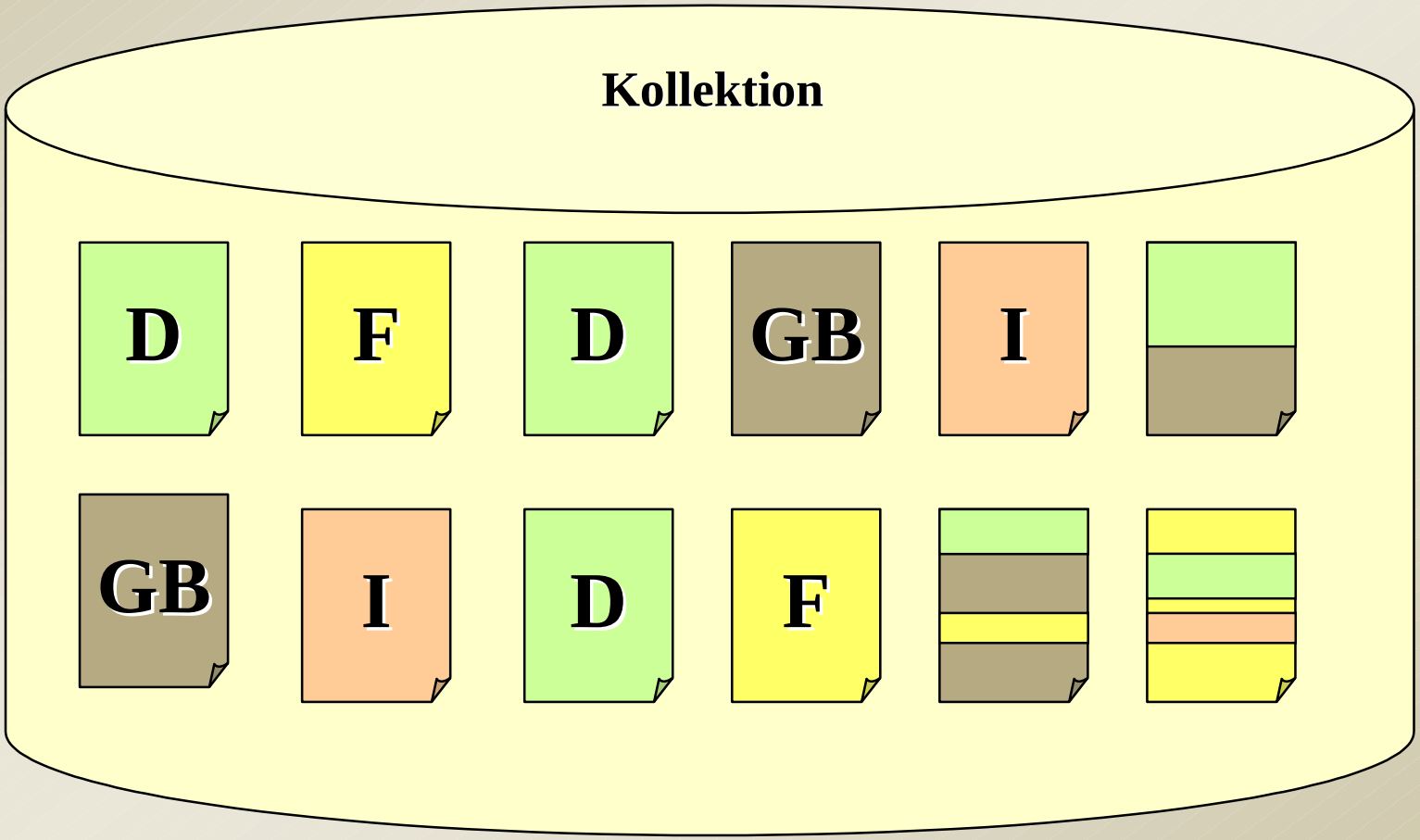



**Fig. 2: Recall-Precision-Graph for Probabilistic DB2 runs with** *DER SPIEGEL*

# Multilinguale Sprachverarbeitung

- Sprachverarbeitung kann die Retrieval-effektivität verbessern ...
  - (Domänenspezifische) Stoppwortlisten
  - Grundformenreduktion / Stemming
  - Kompositazerlegung
  - Mehrwortgruppenanalyse
  - Domänenspezifische Auflösung von Mehrdeutigkeiten
- ...aber sie ist sprachabhängig!

# *Sprachidentifikation*



# Verfahren: Sprachidentifikation

- Auf der Basis von sprachtypischen Sonderzeichen (ä, ü, ß,å, ç, ø)
- Auf der Basis von Funktionswörtern (Artikel, Präpositionen etc.)
- Auf der Basis der Wahrscheinlichkeit des Auftretens von Buchstabenkombinationen (z.B. „sch“ kommt im Französischen nicht vor)
- Auf der Basis der Häufigkeit von Trigrammen der Zeichen (Vgl. mit Trigramm-Datenbank)
- DEMO http://demos.basistech.com/

# Stoppwortlisten

- http://snowball.tartarus.org/german/stop.txt
- http://snowball.tartarus.org/french/stop.txt
- http://www.unine.ch/Info/clef/
- Sinnvolle Erweiterungen um wenig sinntragende Wörter:
  - befassen, Dokument, liefern, Suche, ...
  - find, reports, discuss, comment, mention...
  - contenir, rechercher, concernant, trouver...
  - Informes, relevante, hablar, buscar...

# Stemmer / morph. Analysewerkzeuge

- ZIEL:
  - Simultane Suche von Wortvarianten
  - Indexreduzierung
- Stemmer
  - trennen z.B. Pluralendungen und Genitivendungen ab
- morphologische Analysewerkzeuge
  - Analyse der Wörter mit linguistischen Methoden
  - alle Varianten der Flexion
  - oft in Kombination mit umfangreichen Lexika
- Statistische n-gramm Ansätze

# Stemmer

- z.B. Table look-up Stemmer

| Term | Stem |
|---|---|
| computation<br>Computational<br>Computer<br>Computerization<br>Compute | comput |

# *Kompositaanalyse*

- EN, FR: mehrere tokens
- DE, FI: *ein* token
- Zugänglichmachung über das Kompositum ***und*** über die Bestandteile
  - z.B. Gartentor, Garten, Tor
- manchmal tritt Verschiebung der Bedeutung ein (z.B. Dachstuhl, Feuerzeug)
- Position der Trennung (Glücksautomaten, Mädchenhandelsschule, Industriestrompreise, Nachteilzug, westeuropäische)

# Mehrwortdeskriptoren

- Lexikalisiert:
  - Europäisches Patentamt, Deutscher Bundestag, Angewandte Informatik, Vereinigte Staaten etc.
  - Meist Realisierung über Lexikon
- Nicht-lexikalisierte, komplexe Deskriptoren
  - ADJ-SUB (inkrementelle Meßeinrichtung)
  - GEN (Parkplatz der Universität)
  - KOORD (Haus und Garten)
  - PRÄP (Arbeitskreis IT in Hannover)

# Behandlung von Eigennamen

- Eigennamen bzw. sog. named entities sind sehr gute „Anker“ für das Information Retrieval
- High quality proper name recognizer
- High quality transliteration procedure for named entities
  - München, Munich, Monaco
  - Mailand, Milano, Milan
  - Harry Potter ? Russ. Staatschefs ?
  - Ethnika und ihre Ableitungen (St. Etienne – Stephanois, Hannover - Hannoveraner)

# Sprachvarianten

- American – British English
  - (real) estate manager
- Deutsche Rechtschreibreform
  - Rad fahren, radfahren
  - Prozessakten, Prozeßakten
- Swiss and German German
  - Müllverbrennung – Kehrichtentsorgung
  - Kostenvoranschlag – Offertenstellung
  - German "Rechtschreibreform„

# *Multilinguale Korpora*

- Ableitung von Term-Term-Beziehungen aus vergleichbaren, multilingualen Korpora
- Alignment auf Satzebene / über formale Eigenschaften
- Beispiel:
  - Aligned Hansards of the 36th Parliament of Canada
  - www.isi.edu/natural-language/download/hansard

# *Mehrdeutigkeiten*

1:n customs → Sitten; Zölle

C 162

**<EN-title> EU and Turkish customs</EN-title>**
**<EN-desc> Find documents about the problems posed by Greddce concerning the abolishment of customs restrictions between the European Union and Turkey. </EN-desc>**

# Ausblick

- Erweiterung von MIMOR auf neue Sprachen (Japanisch, Arabisch ...)
- Erweiterung auf Multimedia-Dokumente (Bild- und Audioretrieval)
- Question-Answering
- Evaluierung im CLEF

# *Danke für Ihre Aufmerksamkeit!*