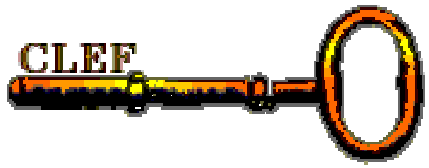
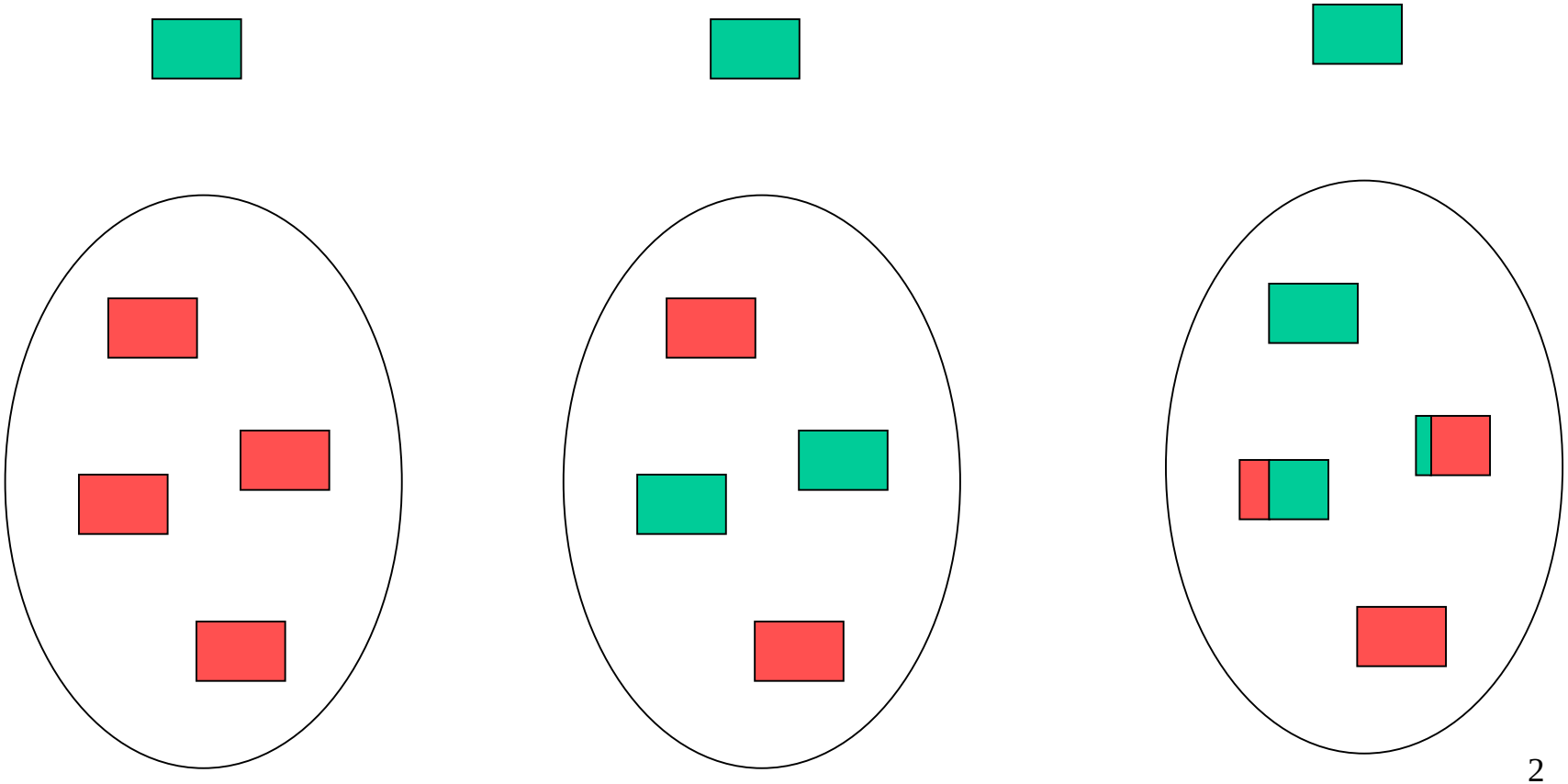


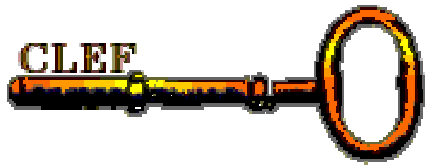
CLEF 2000
State-of-the Art
Multilingual Information Access

Peter Schäuble
Eurospider Information Technology AG
8006 Zürich, Switzerland
schauble@eurospider.com

Multi- & Cross-Lingual Information Access



MLIR Applications

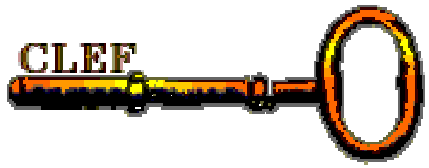


- Multilingual information access in multilingual country, organization, enterprise, etc.
- Cross-language information retrieval for users who read a second language (large passive vocabulary) but are not able to formulate good queries (small active vocabulary).
- Monolingual users may retrieve images by taking advantage of multilingual captions.
- Monolingual users may retrieve documents and have them translated (automatically or manually) in their language.

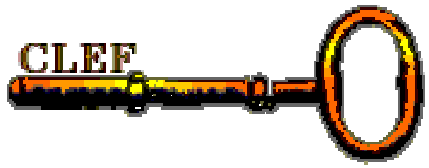
How many Internet users are NOT English speaking?

- **2001 - 49 %**
- **2003 - 54 %**
- **2005 - 59 %**
- **Total number of Internet users will rise from 171 million to 345 million by 2005**

Why is Cross-Language Information Retrieval Important?



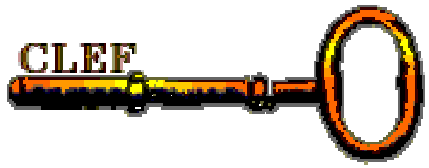
- **More information workers with less time require fast access to global resources**
- **global B2B interactions (virtual enterprises)**
- **global B2C interactions (online trading, travelling)**
- **time critical information (translation comes too late)**



What are the Problems?

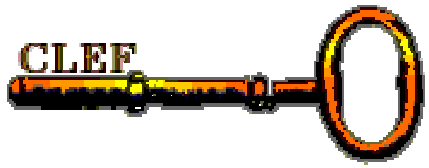
- **Ambiguous terms (e.g. performance)**
- **Multiword phrases may correspond to singleword phrases (e.g. South Africa => Südafrika)**
- **Coverage of the vocabulary**
- **There is not a one-to-one mapping between two languages**
- **Translating queries automatically (lack of syntax)**
- **Translating documents automatically (performance, …)**
- **Computing mixed result lists**

MLIA History



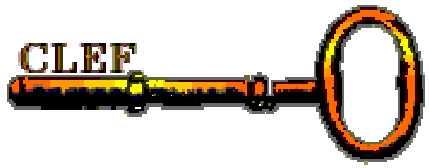
- **1970 Salton runs retrieval experiments with a small English/German dictionary**
- **1972 Pevzner shows for English and Russian that a controlled thesaurus can be used effectively for query term translation**
- **1978 ISO Standard 5964 for developing multilingual thesauri (revised in 1985)**
- **1990 Latent Semantic Indexing (LSI) applied to CLIR**

MLIA History (cont.)



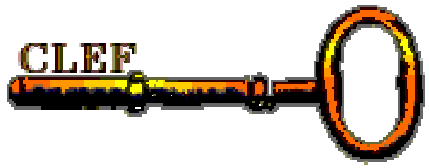
- **1994 1st PhD thesis on CLIR by Khaled Radwan**
- **1996 Similarity thesaurus applied to CLIR (ETH Zurich)**
- **1996 Dictionary based retrieval applied to CLIR (Umass & XEROX Grenoble)**
- **1997 Generalized Vector Space Model (GVSM) applied to CLIR (CMU)**

MLIA History (cont.)



- **1997 CLIR (Cross-Language Information Retrieval) track starts within TREC**
- **1998 NTCIR starts in Japan**
- **1999 TIDES (Translingual Information Detection, Extraction, and Summarization) starts in U.S.**
- **2000 CLEF starts in Europe**

Indexing Multilingual Document Collections



- Stopword Lists
- Tokenizing
- Language Identification
- Normalizing Indexing Features
- Part of Speech Tagging
- Phrase Indexing
- Multiscript Text Processing

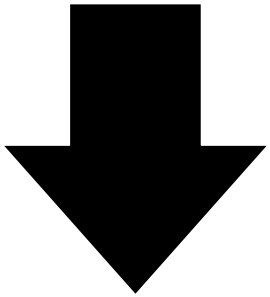
Query

Document

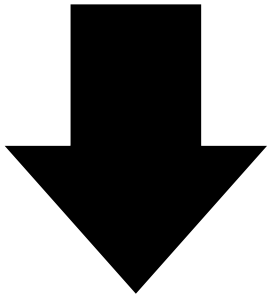
Krebse bekämpfen mit Gift

empoisonner des écrevisses

Da der Rote Sumpfkrebs mit Raubfischen bekämpft werden kann, ist diese Massnahme dem Gifteinsatz gegen Sumpfkrebse vorzuziehen.
L'écrevisse rouge des marais pouvant être combattue par l'introduction de poissons prédateurs, il y a lieu de substituer cette mesure à l'empoisonnement projeté.



Indexing

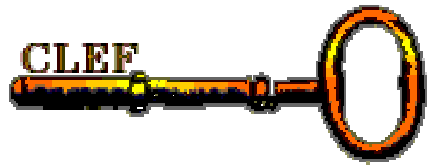


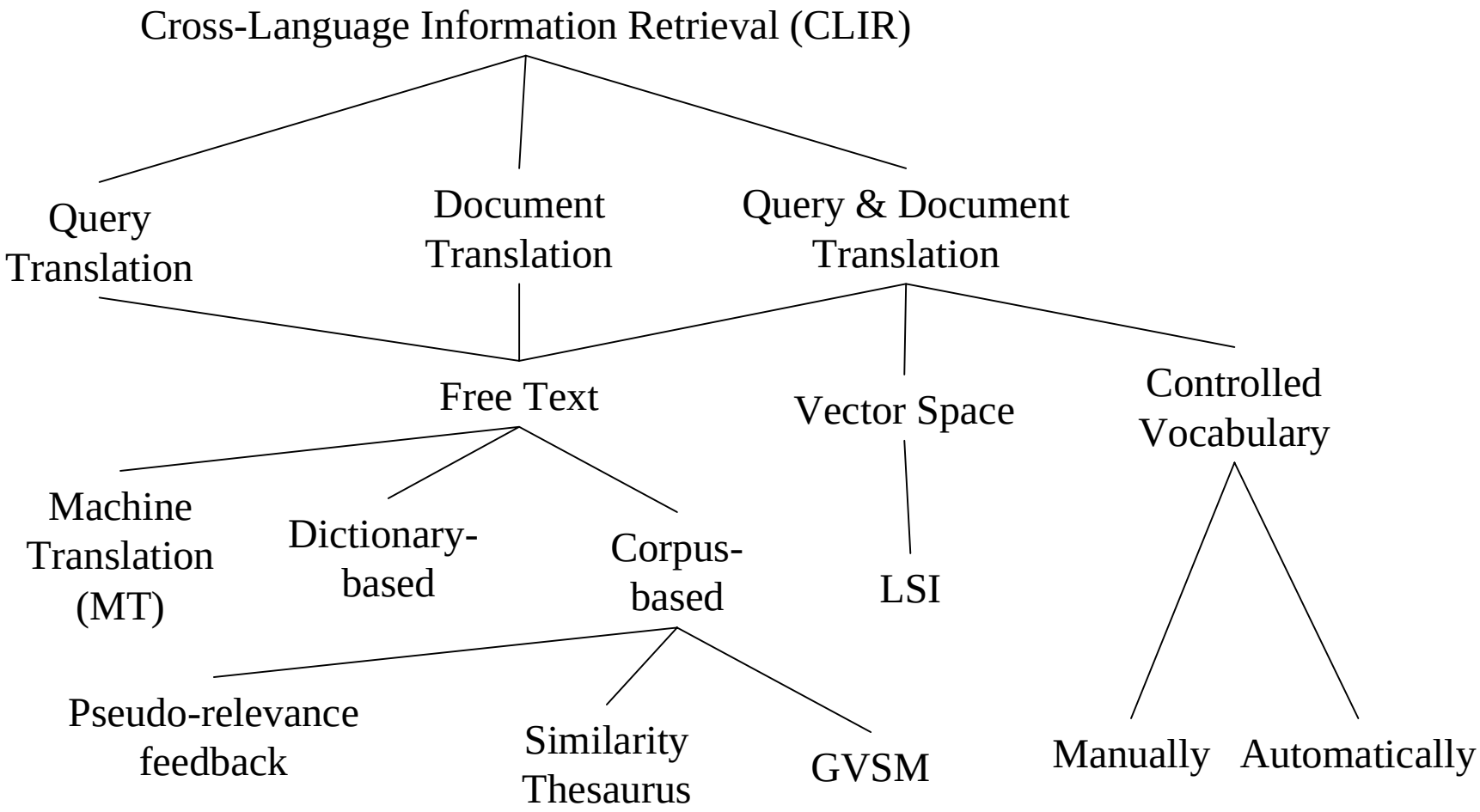
krebs kaempf gift

empoisonner écrevisse

rot sumpf krebs raub fisch kaempf massnehm gift einsetz sumpf krebs vorzieh
écrevisse rouge marais pouvant combattue introduction poisson prédateur lieu substituer mesure empoisonner proje

CLIR Approaches





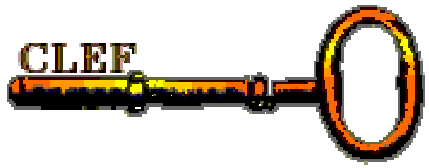
Translating the 400 Million non-English Pages of the WWW

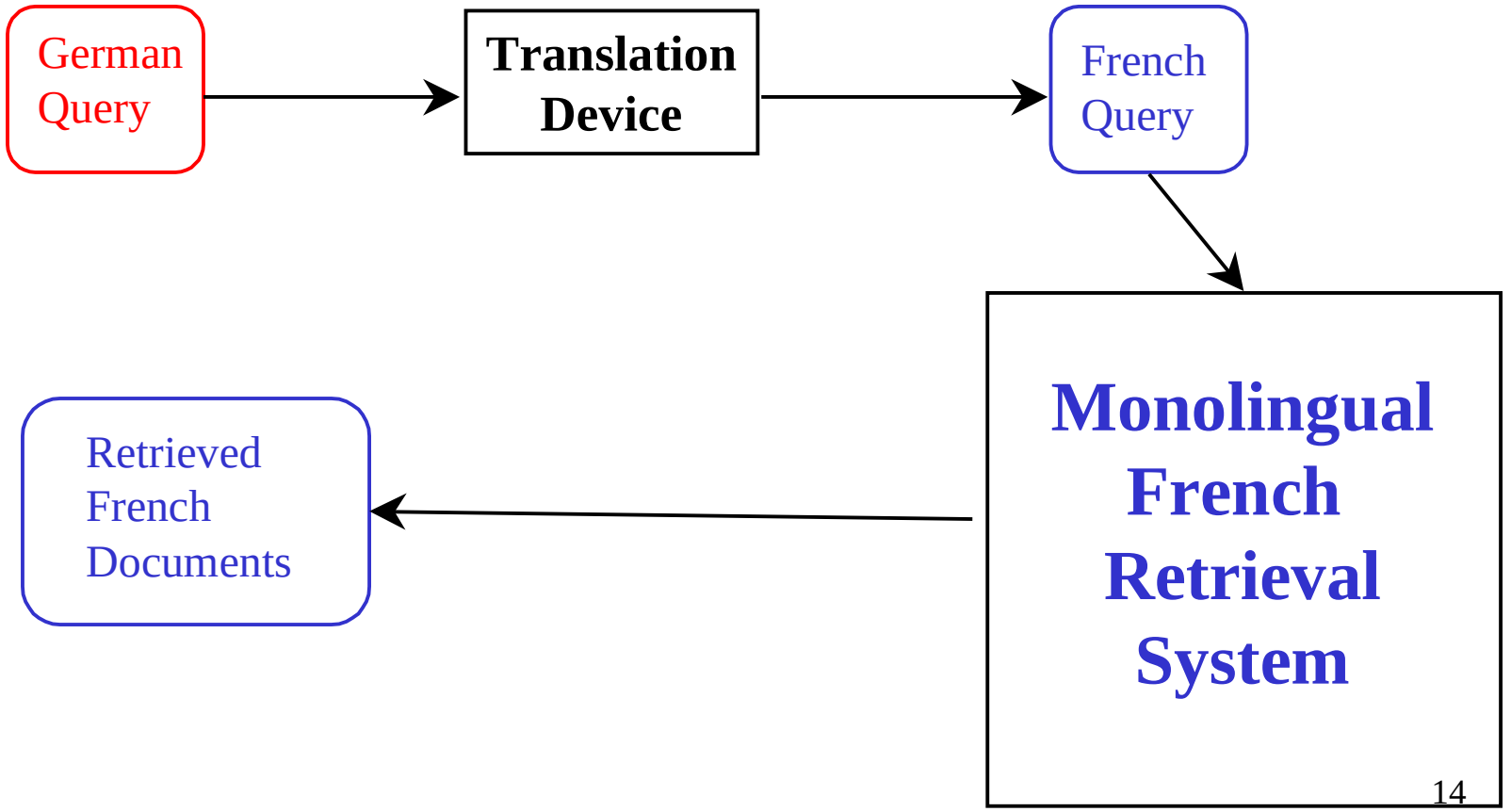
... would take 100’000 days (300 years) on one fast PC.

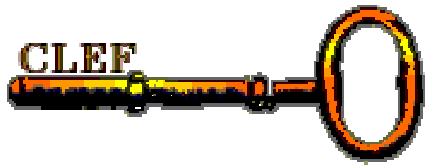
Or, 1 month on 3’600 PC’s.

(L&H- Full MT System)

Query Translation Based CLIR





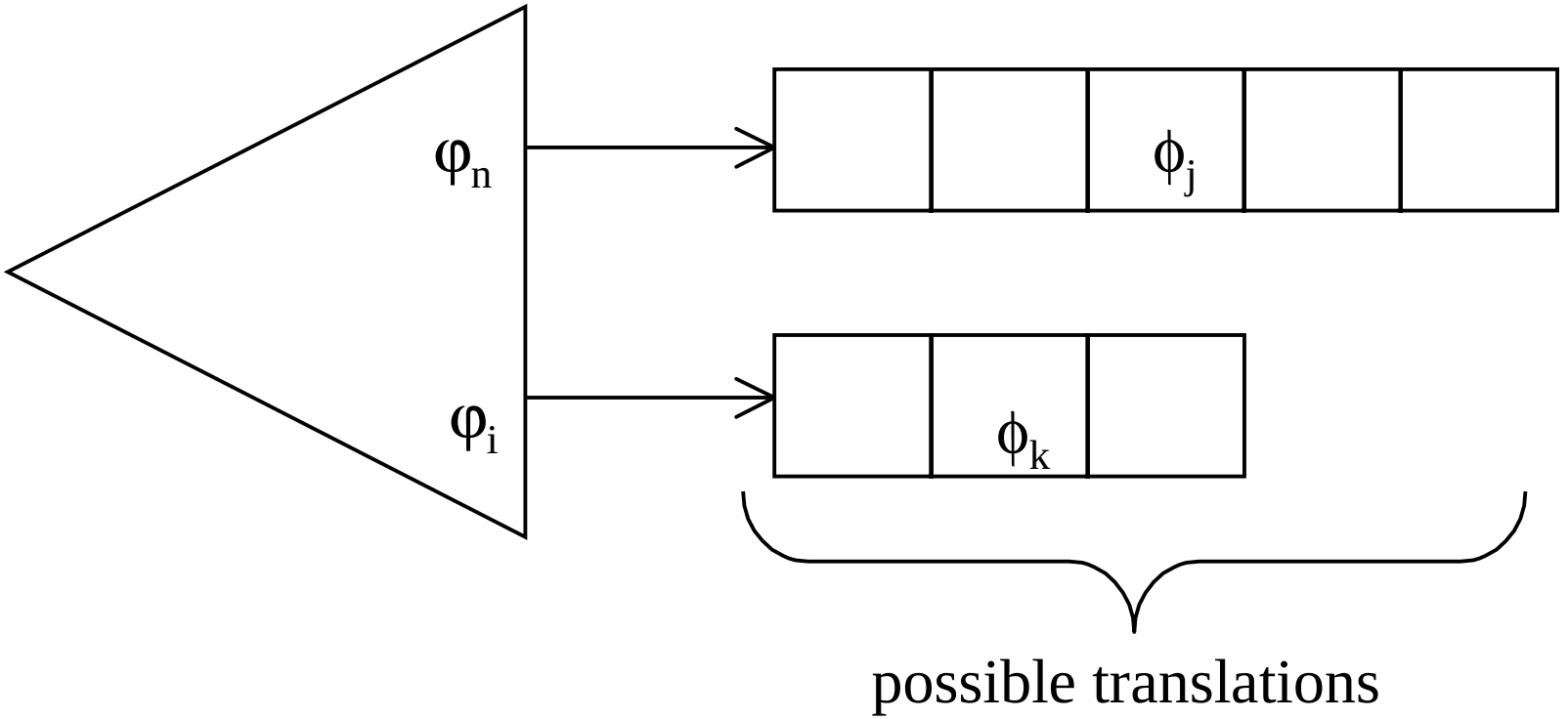


Machine Translation

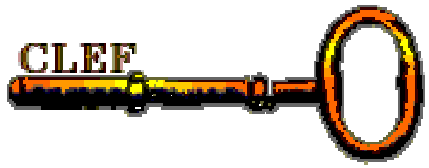
- **Original German Text:**
  Condor-Maschine bei Izmir abgestürzt:
  Mutmasslich 16 Tote.

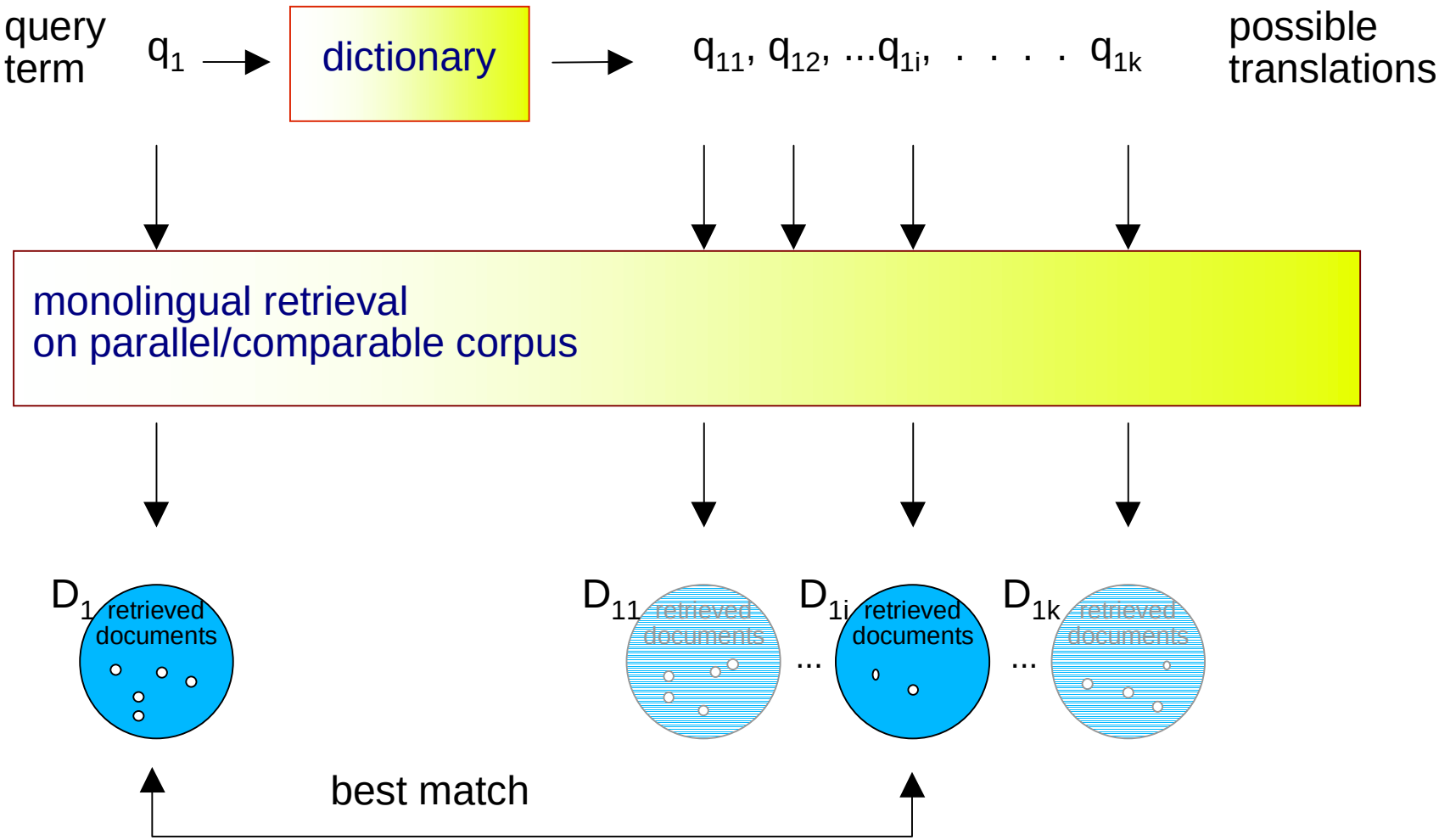
- **English Translation:**
  CONDOR machine with Izmir fallen:
  Courage-isometrically 16 dead ones.

Dictionary Based CLIR

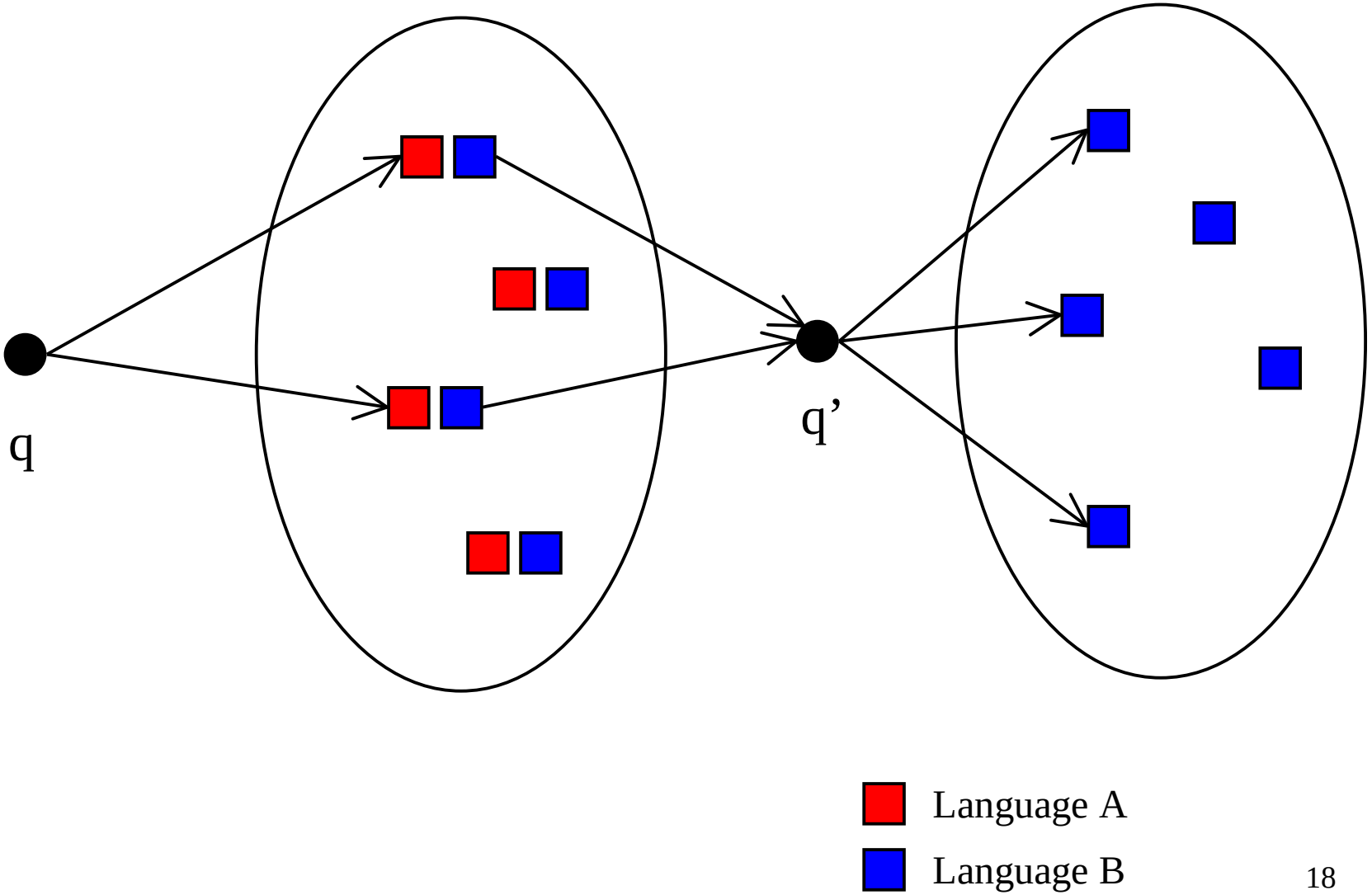


Query Term Disambiguation

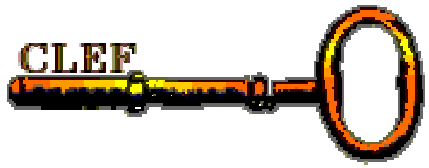




Pseudo-Relevance Feedback illustrated



Generating Mixed Ranked Lists of Documents



- Normalizing scales of relevance
  - using aligned documents
  - using ranks
  - interleaving according to given ratios
- Mapping documents into the same space
  - LSI
  - document translations

CLEF 2000

European Cross-Language Evaluation Forum

- **Sponsored by the DELOS Network of Excellence for Digital Libraries and funded by the Information Societies Technology programme of the European Commission Fifth Framework**
- **20 participating groups (25 registered)**
- **German: Spiegel, Frankfurter Rundschau**
- **English: Los Angeles Times**
- **French: Le Monde**
- **Italian: La Stampa**
- **More information available at http://www.iei.pi.cnr.it/DELOS/CLEF**

CLEF 2000

Cross-Languages Information Retrieval

http://www.iei.pi.cnr.it/DELOS/CLEF

topics
either
E,G,F,I
or…..

documents

English German French Italian

Participant's CLIR
Information Retrieval System

one result list of E, G, F and I
documents ranked in decreasing
order of their estimated relevance