Language Technology

Multilingual and Crosslingual Information Retrieval and Access

Feiyu Xu
DFKI, LT-Lab
Germany

Language Technology

Multilingual Information System

- Motivation
- Strategies
- MIETTA System

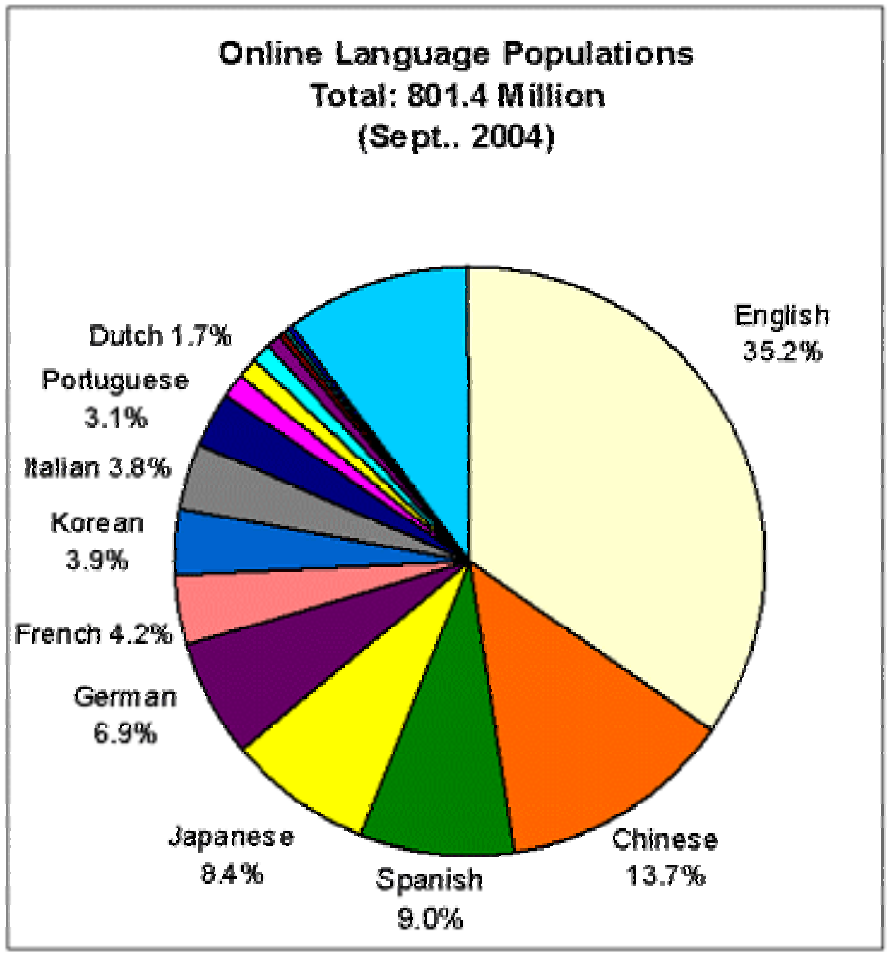
Motivation

- Societal benefits
  - Information exchange to improve understanding
- Economic benefits
  - Information to provide competitive advantage
- Crisis response
  - Language differences can produce costly delays

Source: Douglas W. Oard, IRAL99

More and more web information are encoded in other languages than English, for example, Chinese 13.7%

English is loosing its dominance



Source: http://www.global-reach.biz/globstats/index.php3

Organized Research and Development Activities

Language Technology

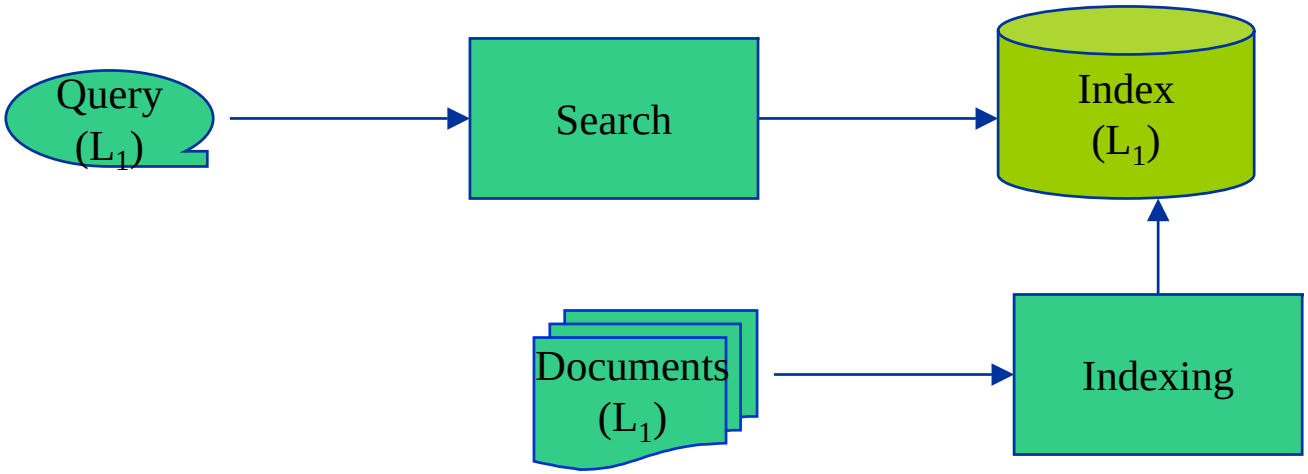
- Text REtrieval Conference (TREC) (http://trec.nist.gov/)
  - Arabic, English, Spanish, Chinese, etc.
  - TREC: crosslingual information retrieval: http://www.glue.umd.edu/~dlrg/clir/trec2002/
- Cross-Language Evaluation Forum (CLEF):
  - http://www.clef-campaign.org/
- NTCIR (NII-NACSIS Test Collection for IR Systems) workshops:
  - http://research.nii.ac.jp/ntcir/workshop/
- Information Retrieval for Asian Language Conference (IRAL)
- European ESPRIT consortium (French, Belgian, German)

What is Information Retrieval (http://www.lt-world.org)

- **Synonyms**: document retrieval
- **Definition**: Information Retrieval is the process of locating information that fits a user's requirements, where the requirements are usually expressed as a search query. The fit of the retrieved information with the information need is referred to as "relevance“ …
- http://www.lt-world.org/HLT_Survey/ltw-chapter7-2.pdf

What is Monolingual Information Retrieval?

- Query and information to be looked for are encoded in a same language



What is Multilingual Information Retrieval?

- An extension of the general information retrieval problem
- Finding information, e.g., web documents which are not encoded in the same language as the query is encoded in
- Similar terms: “crosslingual information retrieval” and “translingual information retrieval”

Allow anyone to find information that is expressed in any language

يا ليلي يا عيني

Исследований

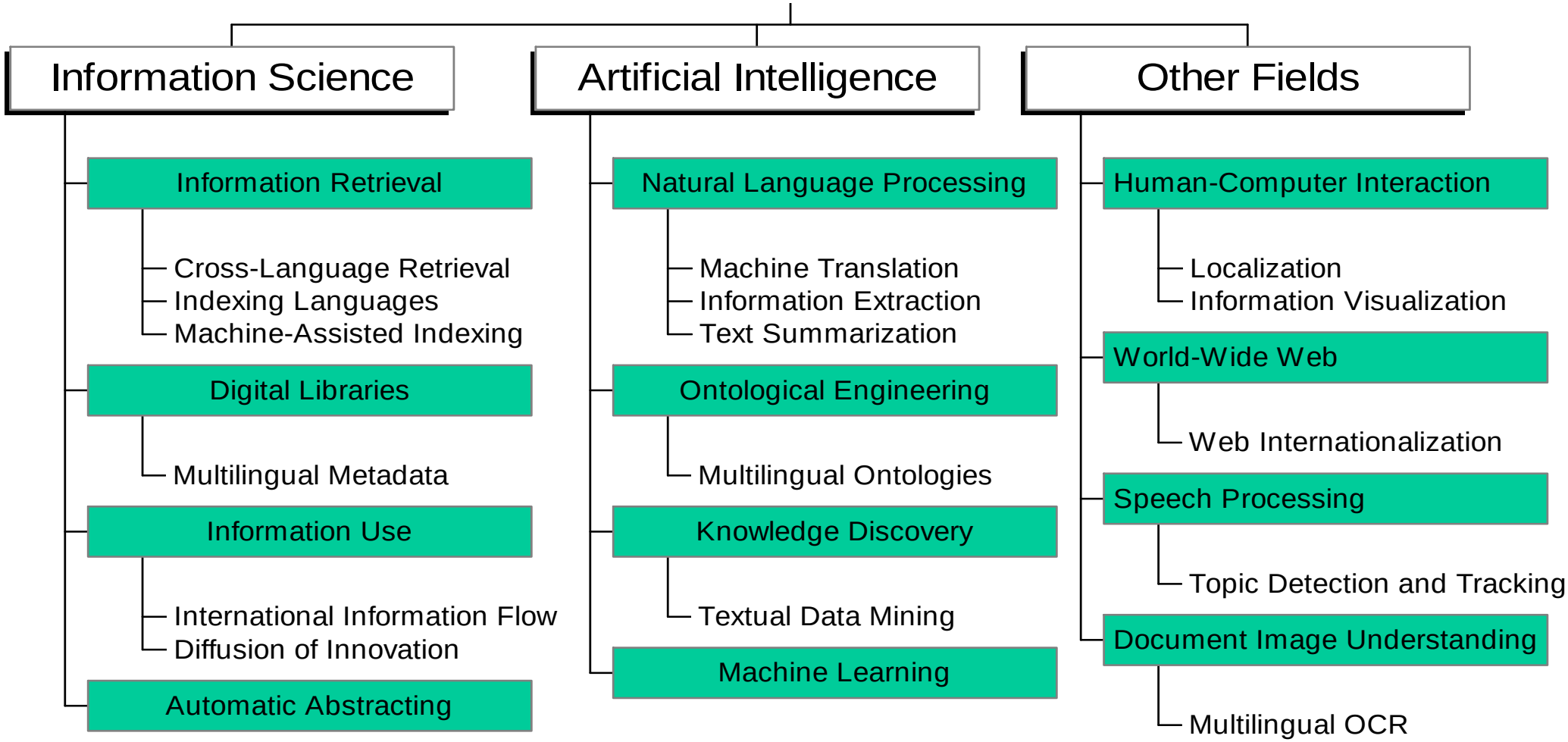
高等学校

att förstå

których można

Source: Douglas W. Oard, IRAL99

Multilingual Information Access



Source: Douglas W. Oard, IRAL99

Different Multilingual Information Retrieval Strategies Supported by Language Technologies

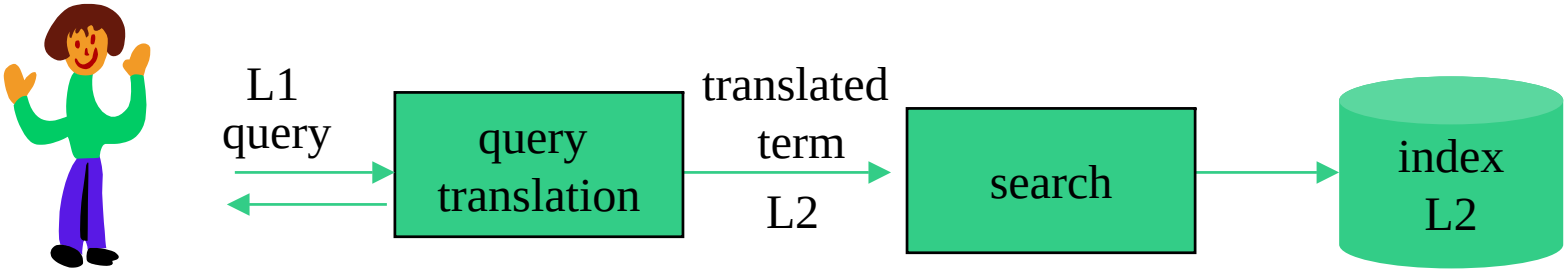
Language Technology

- Online query translation
  - Help user to formulate his query in a foreign language
- Online document translation
  - Translate the found document into the query language
- Offline document translation
  - Make web documents multilingual available
- Combination of information extraction and multilingual generation
  - Make database information multilingual available and allow the free text retrieval of database information

Query Translation

Language Technology

- Help user to formulate their query in another language



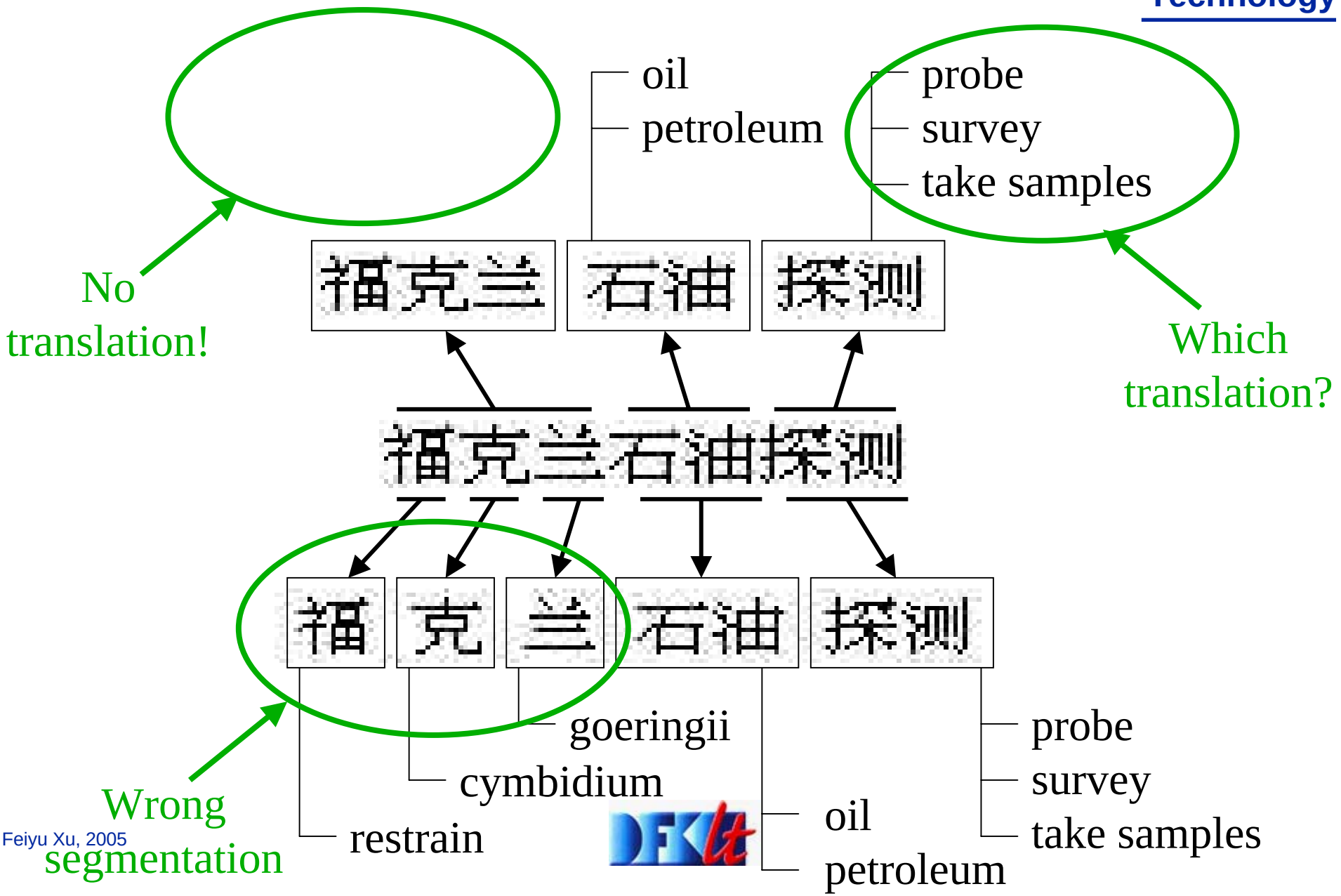
- The primary problem is that short queries provide less context for word sense disambiguation, and inaccurate translations lead to bad recall and precision

- How can the user access the content of the found document?

Three Key Challenges

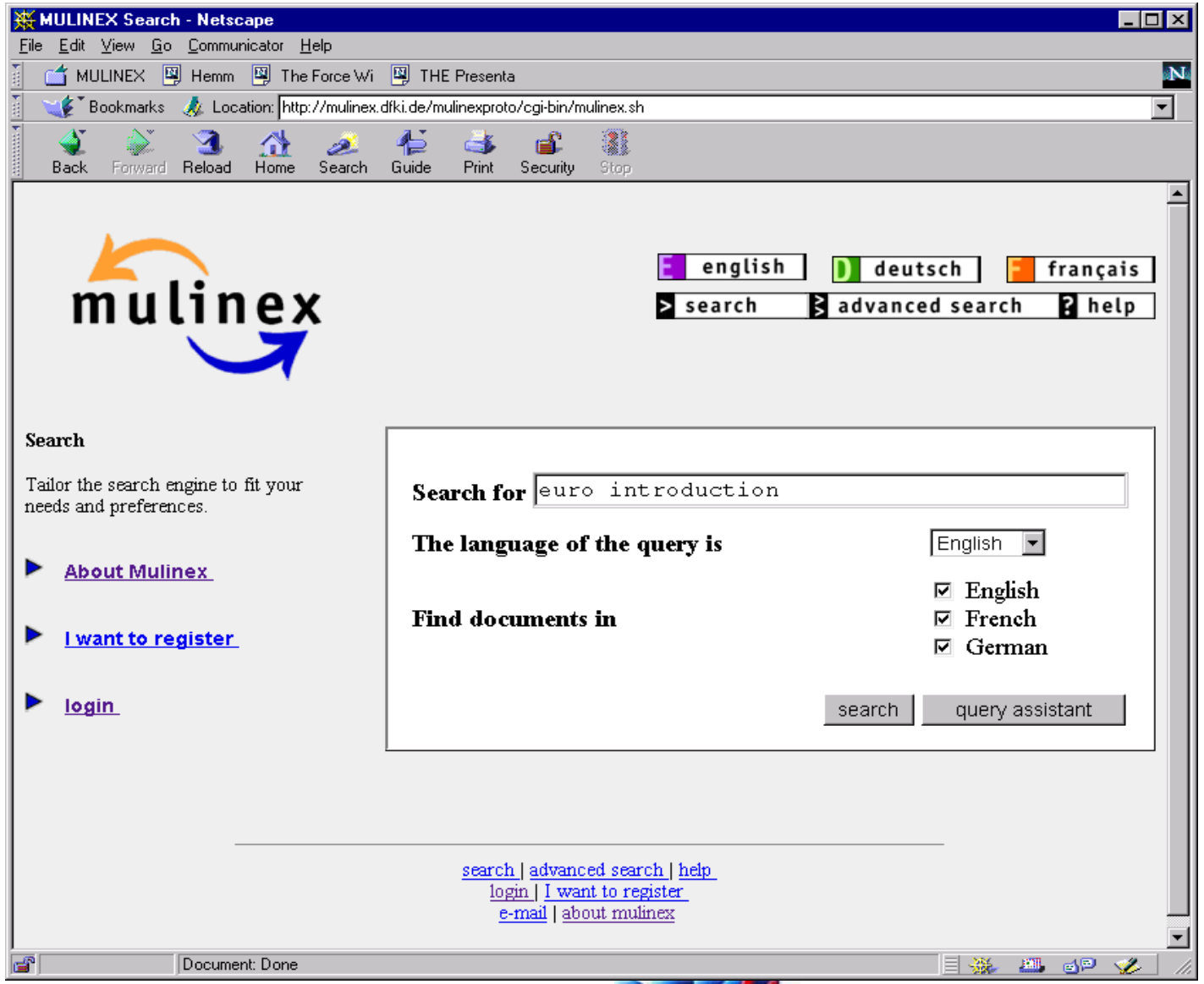
Source: Douglas W. Oard, IRAL99

Language Technology



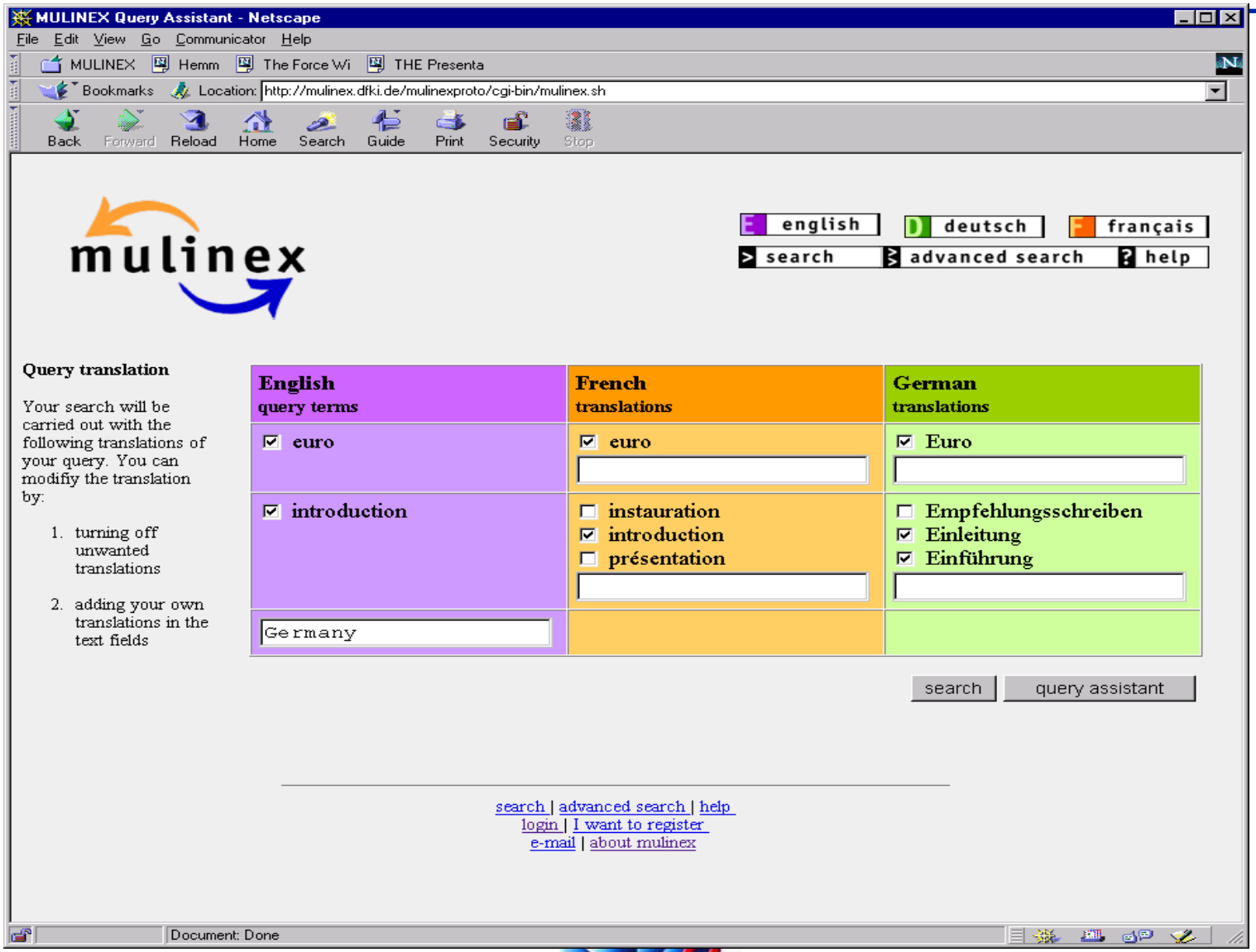
MULINEX System

Language Technology



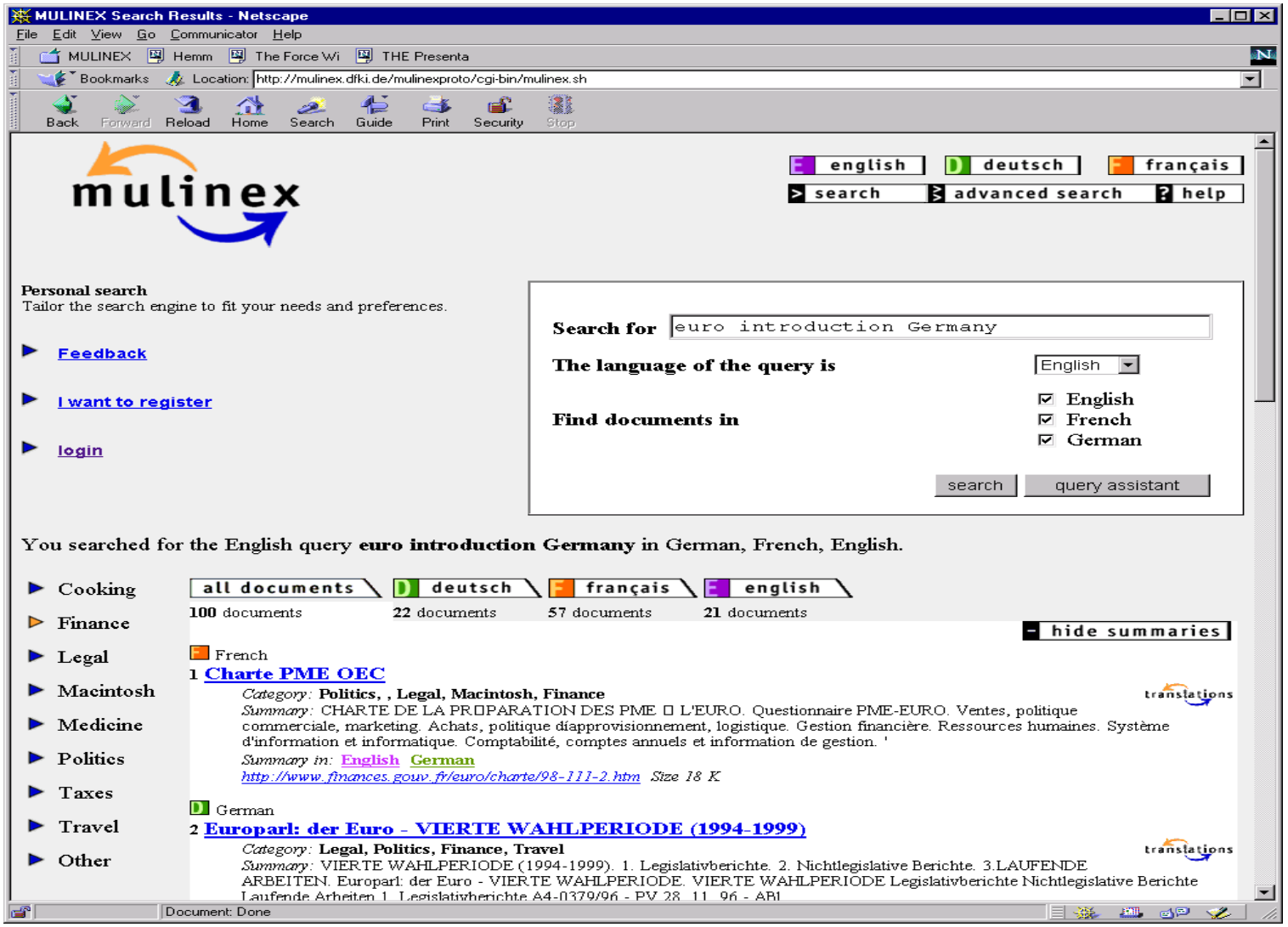
MULINEX System

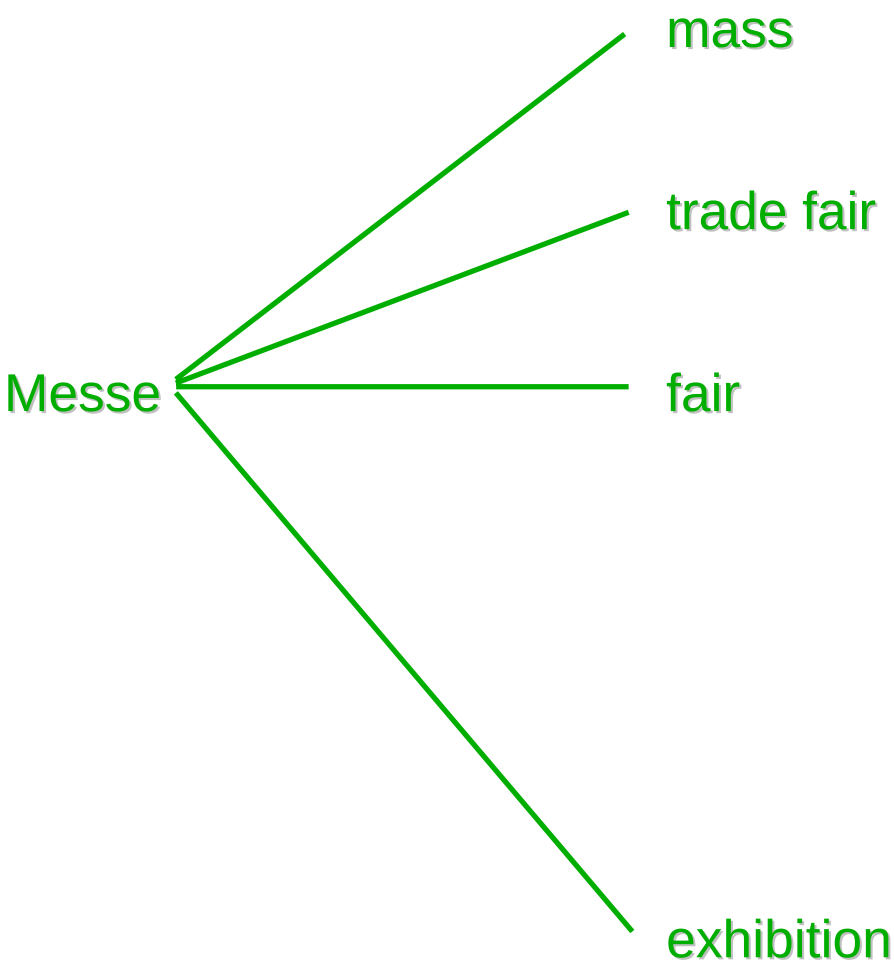
Language Technology

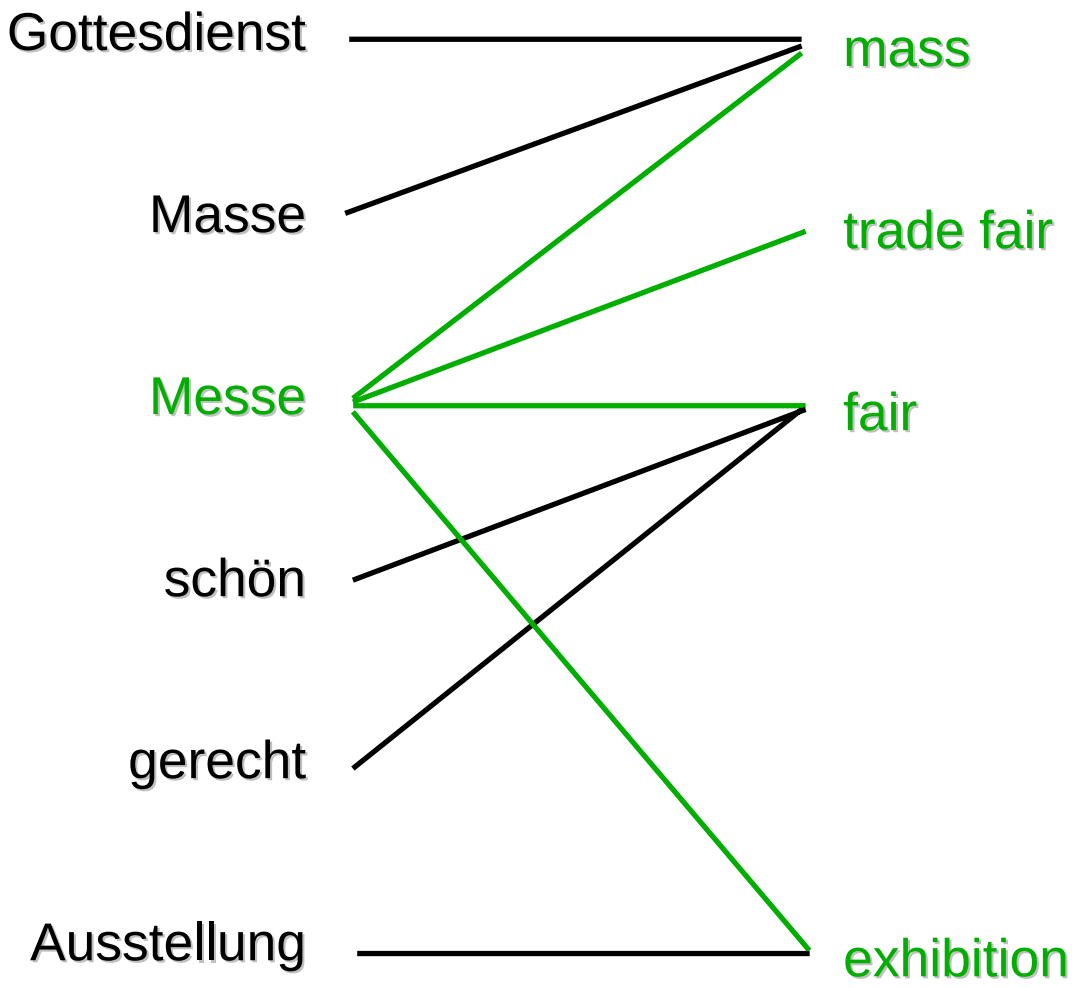


MULINEX System

Language Technology

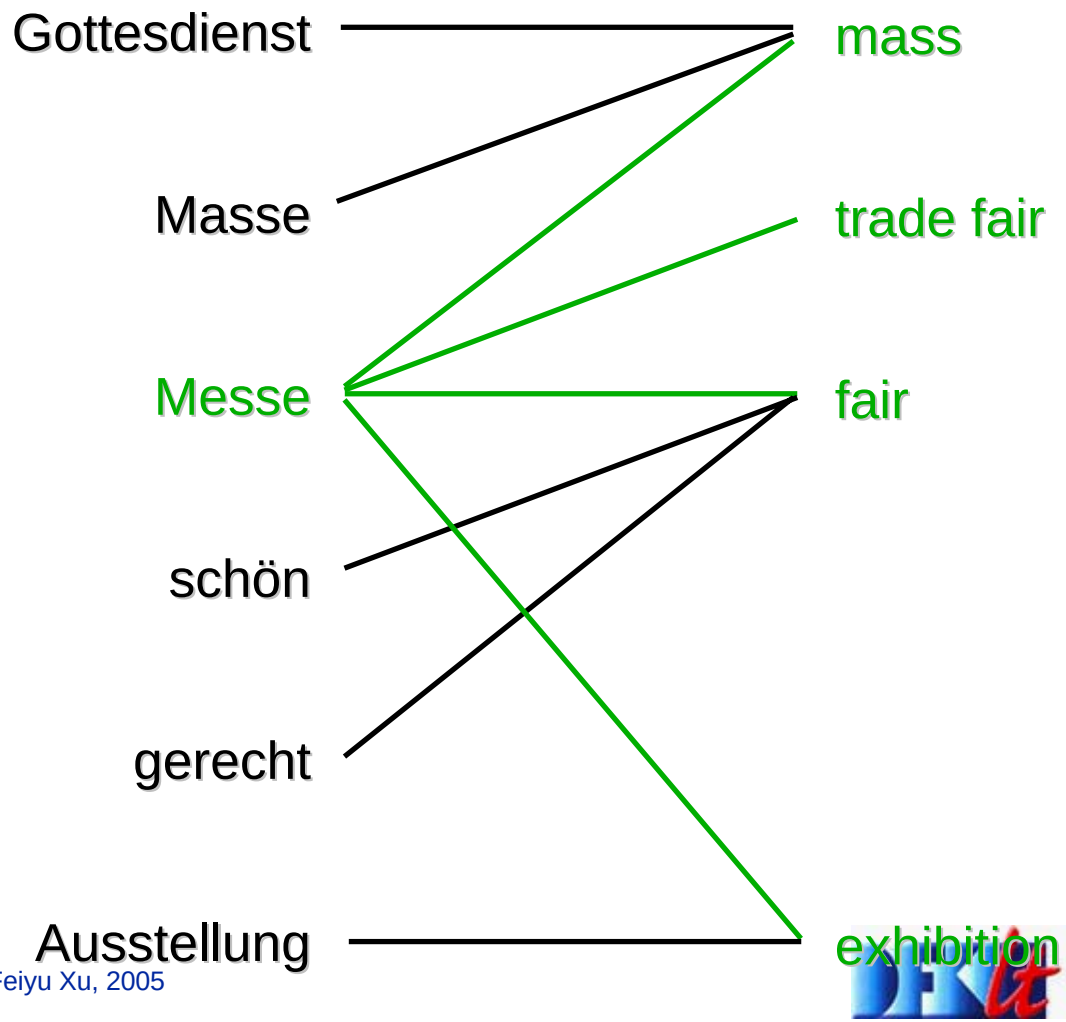



mass
trade fair
Messe
fair
exhibition


Gottesdienst
Masse
Messe
schön
gerecht
Ausstellung
mass
trade fair
fair
exhibition

EXAMPLE

Language Technology



Messe, Gottesdienst, Masse, → mass

Messe → trade fair

gerecht, schön, Messe → fair

Ausstellung, Messe → exhibition

USER FEEDBACK

Language Technology

Messe, Gottesdienst, Masse	→ mass	☑
Messe	→ trade fair	☑
gerecht, schön, Messe	→ fair	☐
Ausstellung, Messe	→ exhibition	☐

MuchMore

Language Technology

Project Goals

Application (MuchMore Demo)

⇒ Addressing a Real-Life Medical Scenario for Cross-Lingual Information Retrieval (CLIR)

Research & Development

⇒ Developing Novel, Hybrid (Corpus-/Concept-Based) Methods for Handling this Scenario

Evaluation

⇒ Evaluating the Technical Performance of (Combinations of) Existing and Novel Methods

Source: I²R, Singapore: January 15th,2003, Paul Buitelaar

MuchMore

Language Technology

Project Partners

CSLI	Stanford University, USA
DFKI	Saarbrücken, Germany
EIT	Zürich, Switzerland
LTI	Carnegie Mellon University, USA
XRCE	Grenoble, France
Zinfo	Frankfurt, Germany

MuchMore

Language Technology

R&D Topics

Annotation-Based CLIR

⇒ Term Tagging (incl. Disambiguation)

⇒ Relation Tagging (incl. Filtering, Discovery)

Classification-Based CLIR

Multi-Document Summarization

Term Tagging

Language Technology

Semantic Resources

Medical Domain

UMLS: Unified Medical Language System

Medical MetaThesaurus (only MeSH2001 is used)

English, German, Spanish, …
730.000 Concepts
9 Relations (Broader, Narrower,…)

Semantic Network

134 Semantic Types
54 Semantic Relations

General

WordNet (EN), GermaNet (DE), EuroWordNet (“linked”)

Term Tagging

Language Technology

Semantic Resources (UMLS)

Concept Names (MRCON): 1.734,706		
ENGLISH 1.462,202	GERMAN 66,381	other languages

C0019682|ENG|P|L0019682|PF|S0048631|HIV|0|
C0019682|ENG|S|L0020103|PF|S0049688|HTLV-III|0|
C0019682|ENG|S|L0020128|VS|S0049756|Human Immunodeficiency Virus|0|
C0019682|ENG|S|L0020128|VWS|S0098727|Virus, Human Immunodeficiency|0|
C0019682|FRE|P|L0168651|PF|S0233132|HIV|3|
C0019682|FRE|S|L0206547|PF|S0277133|VIRUS IMMUNODEFICIENCE HUMAINE|3|
C0019682|GER|P|L0413854|PF|S0538136|HIV|3|
C0019682|GER|S|L1261793|PF|S1503739|Humanes T-Zell-lymphotropes Virus Typ III|3|

Each CUI (Concept Unique Identifier) is mapped to one out of 134 semantic types or TUI (Type Unique Identifier)

Clozapine : C0009079 → *Pharmacologic Substance* : T121

Semantic Types are organized in a Network through 54 Relations

T121|T154|T047

Term Tagging

Language Technology

Semantic Resources (EuroWordNet)

Synonyms between Languages (i.e. German, English, etc.) are Linked Through a Common Interlingual Index (ILI) Code

ILI Code	SynsetID	Synset
3824895	DE-0405065	Fingergelenk, Fingerknochen
3824895	DE-4848521	Knöchel
3824895	EN-2394238	knuckle, knuckle joint, metacarpophalangeal joint

German	7.829 Nouns	2.997 Verbs
English	60.521 Nouns	11.363 Verbs

GermaNet (Used in Development)

⇒ German ~ 25.000 Nouns, ~ 6.000 Verbs, ~ 3.500 Adjectives

Term Tagging

Language Technology

Sense Disambiguation (Methods)

Domain Specific Sense

⇒ Concept Relevance in Domain Corpus

Mineral *0.030774033: Mineralstoff, Eisen, Ferrum, Fluor, Kalzium, Magnesium*
4.9409806E-5: Allanit, Alumogel, ..., Axionit, Beryll, ... Wurtzit, Zirkon

Instance-Based Learning

⇒ Unsupervised Context Models (n-grams)

Training (Learn Class Models)

He drank <milk LIQUID>
He drank <coffee LIQUID>
He drank <tea LIQUID>
He drank <chocolate FOOD, LIQUID>

Application (Apply Class Models)

He drank <chocolate FOOD, LIQUID>
He drank <Java GEOGRAPHICAL, LIQUID>

Reference

Dominic Widdows, Stanley Peters, Scott Cederberg, Chiu-Ki Chan, Diana Steffen, Paul Buitelaar

Unsupervised Monolingual and Bilingual Word-Sense Disambiguation of Medical Documents using UMLS

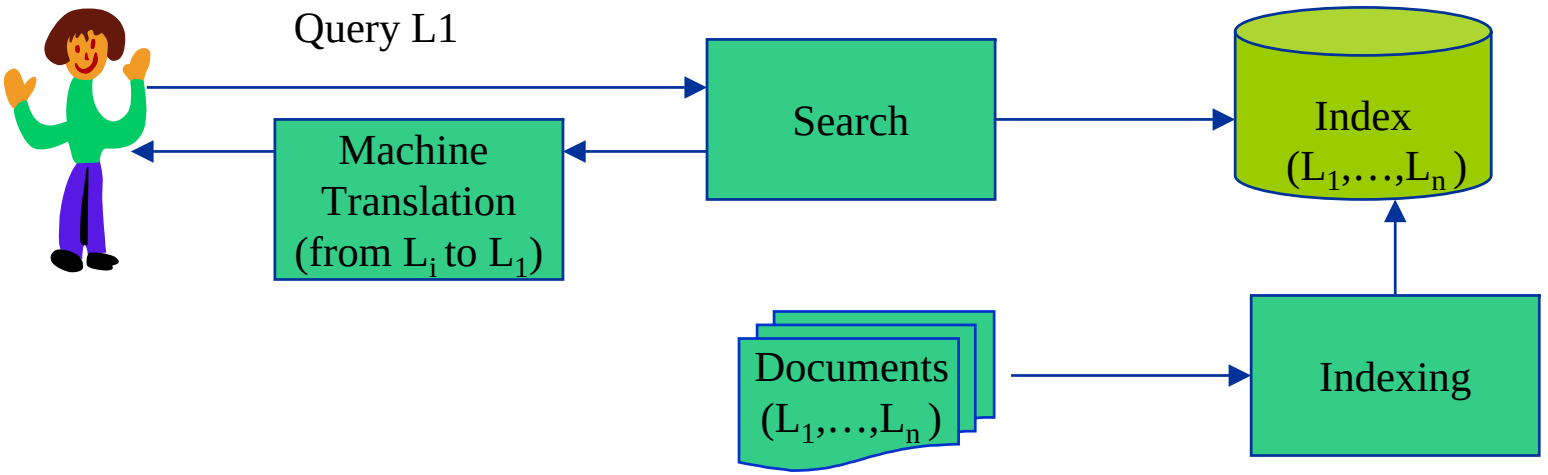
In: **Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine**, Sapporo, Japan, July 11th, 2003

http://dfki.de/~paulb/biomed-wsd.pdf

Online document translation

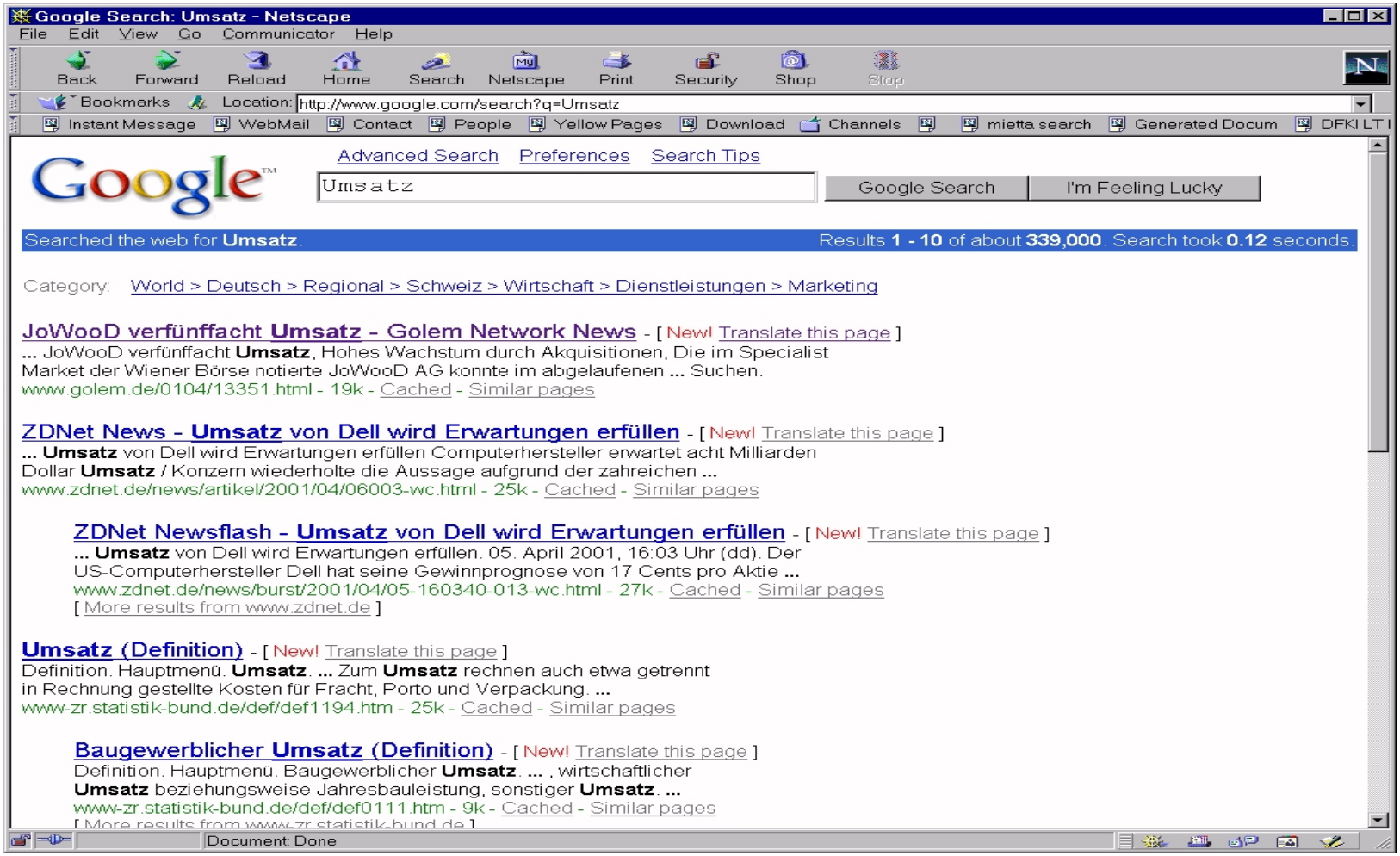
Language Technology

- Translating the found the documents into query language, for example, google



Online document translation (Google)

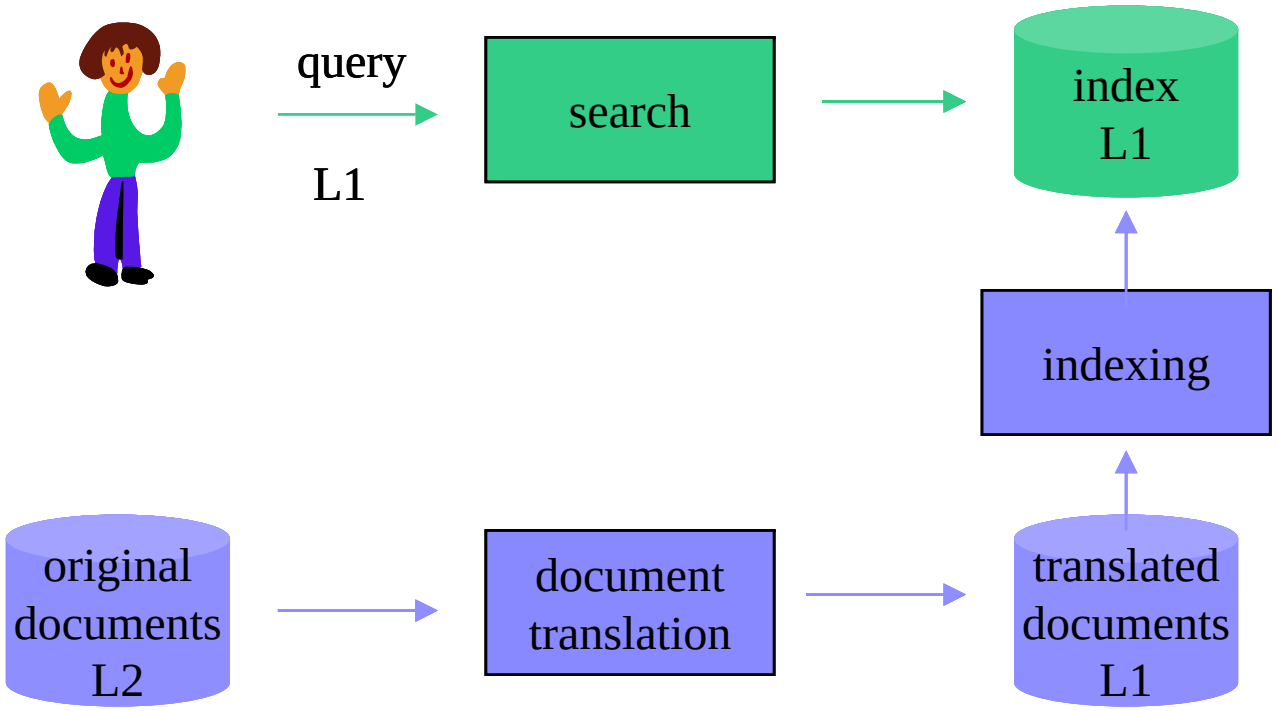
Language Technology



Offline Document Translation

Language Technology

- Automatic offline translation
  - Source text is translated into target languages
  - Index is constructed from translation
  - Search term in one language yields original and translated documents



Offline Document Translation

Language Technology

- A higher translation and retrieval performance, since the full original document provides more context for disambiguation. The word sense disambiguration problem is less severe than query translation
- The main limitation is the duplication of the indices, and the translated documents also need to be stored
- The offline translation is practically not viable due to big cost of computation and storage for the general search engines like Alta-Vista, Yahoo, etc.

Language Technology

Facts Sheet - MIETTA

- Title: MIETTA -Multilingual Information Extraction for Tourism and Travel Assistance
- Funding: EU Language Engineering Sector of TAP (HLT-IST)
- Technical Partners: DFKI, Celi, University of Helsinki, Polito, Unidata
- User Partners: Commune DI Rome, City of Turku, Staatskanzlei of the Saarland

Objectives

Language Technology

- Multilingual internet portal and specialised information system for tourist information

  Five languages: English, Finnish, French, German, Italian

  Three regions: Rome, Saarland and Turku

- Integrated access to heterogeneous data sources and make it fully transparent to end users whether they are searching in
  - WWW documents or
  - Databases

Offline Document Translation in MIETTA

Language Technology

- Use document translation as the main strategy. The reason is that it allows direct access to the content, it provides better performance within a restricted domain

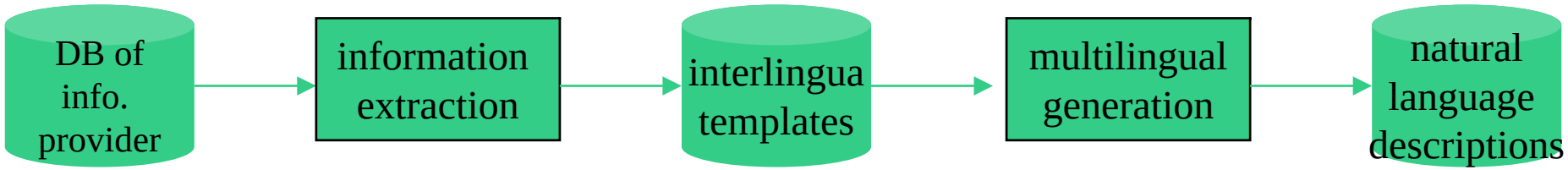
- Use LOGOS for document translation, which covers the following directions:
  - German⇒ English, French, Italian
  - English⇒ French, German, Italian, Spanish

- The final document collection in MIETTA after the document translation yielded an almost fully covered multilingual setup.

Information Extraction and Multilingual Generation

Language Technology

- Motivation
  - Make the database content more structured and multilingual accessible.
  - Apply the same free text retrieval method to the generated descriptions as to the web documents



Information Extraction

Language Technology

- The objective of information extraction is twofold:
  - To extract the domain relevant information (templates) from the unstructured data so that the user can access more facts and more accurately
  - To normalise the extracted data in a language independent format to facilitate the multilingual generation

- Three steps for template extraction in MIETTA
  - Natural language shallow processing: named entities, np, vp
  - Normalisation: converting information into a language independent format
  - Template filling: mapping the extracted information into template by employing specific template filler rules

Example of IE

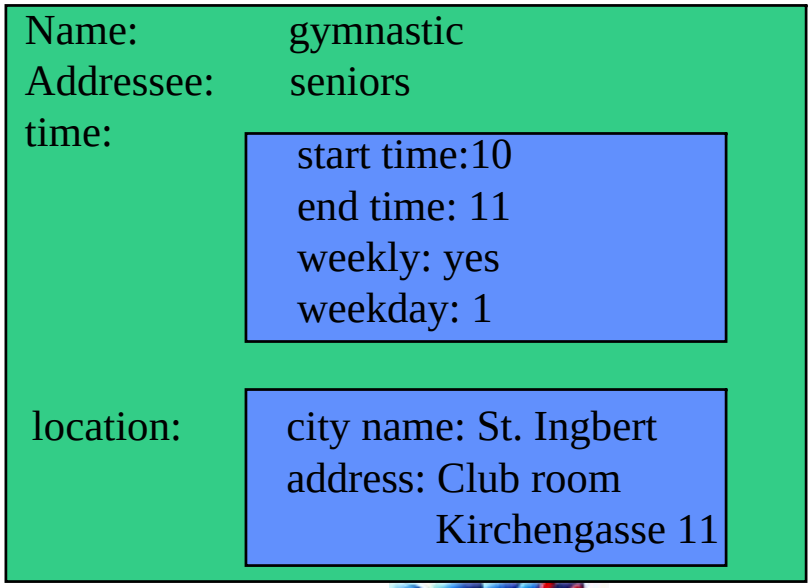
Language Technology

German text from an event calendar in Saarland

St. Ingbert: -Sanfte Gymnastik für Seniorinnen und Senioren, montags von 10 bis 11 Uhr im Clubraum, Kirchengasse 11.

English: St. Ingbert: -Gentle Gymnastic for seniors, every Monday from 10:00 to 11:00 am, in Club room, Kirchengasse 11

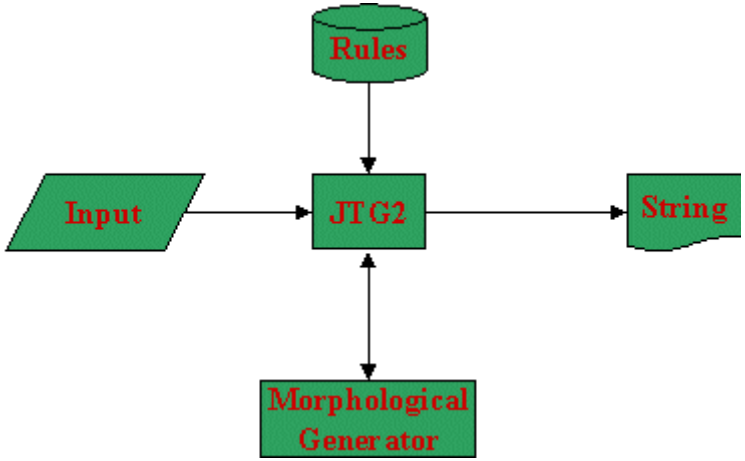
Event:



Multilingual Generation

Language Technology

- Template Generation system (JTG/2)



- Language independent input allows for easy extension of the generation component to other languages

Example

Language Technology

Level1: Event
Level2: Theater
Level3:
Event-Name: Faust
StartDate: 21.10.99
PlaceName: Staatstheater
Address: Schillerplatz, 66111 Saarbrücken
Phone: 0681-32204

English:

The theater show Faust will take place at the Staatstheater in Schillerplatz 1, 66111 Saarbrücken (in the downtown area).

The scheduled date is Thursday, October 21, 1999. Phone: 06 81-32204

Finnish:

Teatteriesitys Faust järjestetään Staatstheaterissa, osoitteessa Schillerplatz 1, 66111 Saarbrücken (keskustan alueella). Tapahtuman päivämäärä on 21. lokakuuta 1999. Puhelin: 06 81-32204.

Mietta Framework

Language Technology

Query L1

Query Translation

Query L2

Form-based Query

Interlingual Templates

Index L1

Free Text Query

Index L2

Information Extraction

Multilingual Generation

Data Base

Document Base L1

Document Translation

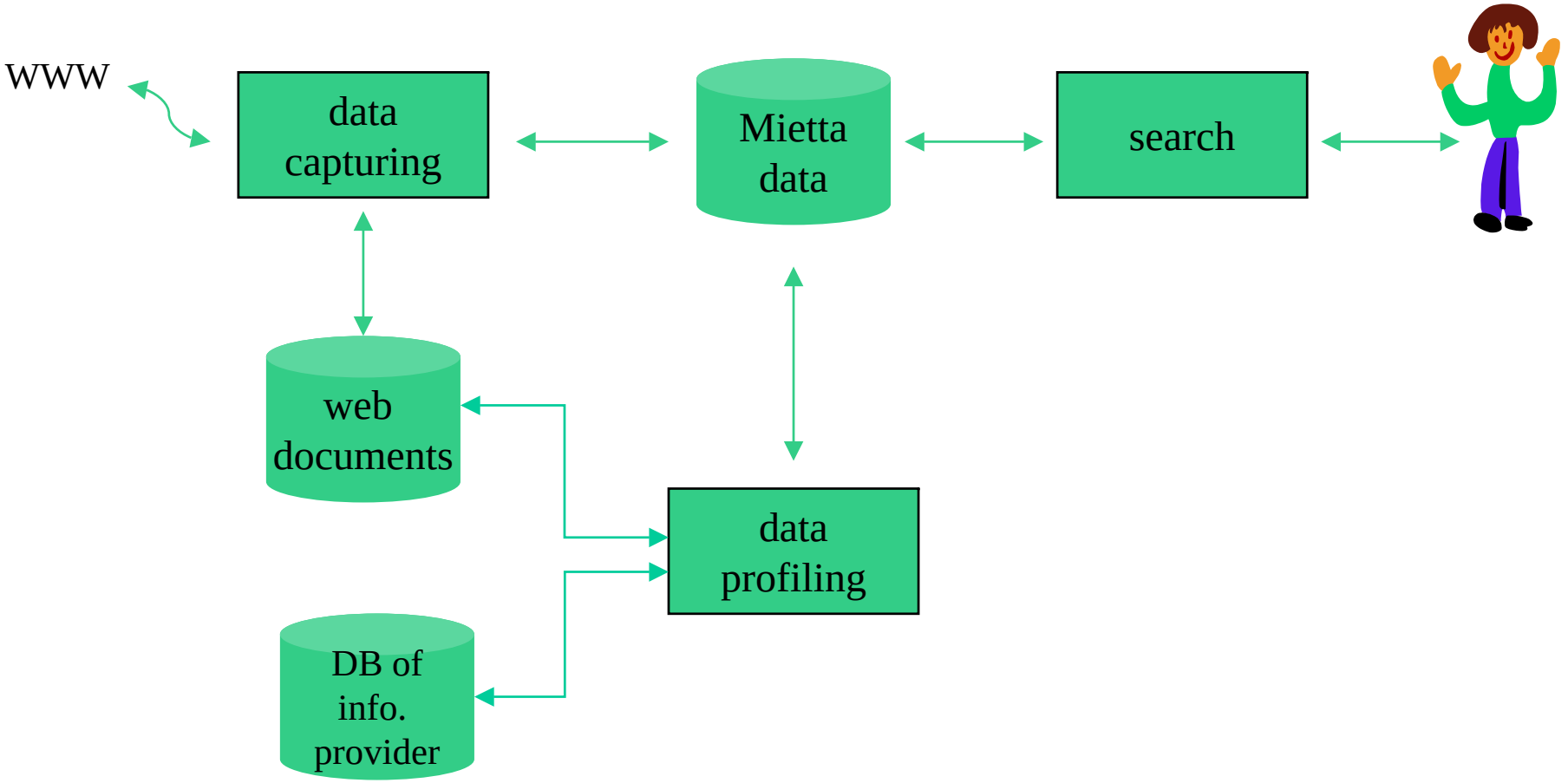
Document Base L2

DFKI

Feiyu Xu, 2005

The Overall MIETTA System

Language Technology



Data Profiling

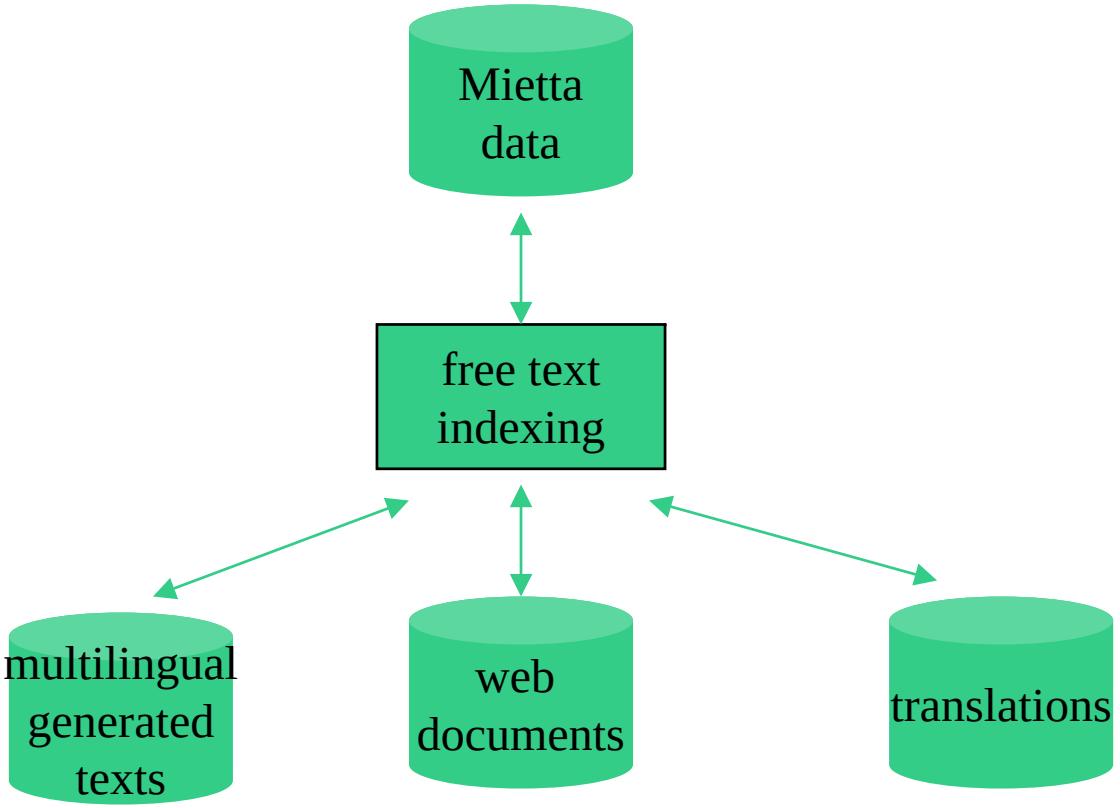
Language Technology

- Document translation, based on LOGOS machine translation system
- Information Extraction from database entries for template construction
- Multilingual generation from templates to obtain natural language descriptions
- Free text indexing

TNO (ISM, VSM) Indexing Toolkit

Language Technology

- ISM: A lemma-based fuzzy index based on trigrams
- VSM: A vector space model index based on lemmatas



Scalability of the Framework

Language Technology

- Adaptation to other domains
  - Domain specific templates
  - Domain Concept hierarchy
  - Domain specific template filler rules
  - Domain specific generation grammars
- Extension to other languages
  - Natural language generation tool requires less effort for the development of a grammar rule set in a language
  - Information extraction requires available language specific resources
  - Document translation is dependent on the machine translation system

Evaluation of the MIETTA System

Language Technology

- The standard relevance assessment model used in ad hoc and routing forums of TREC is difficult to apply to the complete MIETTA system because of
  - Broad variety of search strategies
  - Heterogenous data sources
- MIETTA is evaluated as technically “excellent” by EU
- Two projects are derived from MIETTA
  - Natural science foundation of China Project in SJTU
  - EU project of MIETTA to transfer the idea into product in XtraMind in Saarbrücken

Conclusion: Innovative Technical Features

Language Technology

- Integration of different multilingual and crosslingual search technologies
- Combination of IE and multilingual generation
- Integration of DB and text document access
- Intelligent User Interface
- XML for advanced information management
- Localisation technologies for user interface and multilingual generation
- Highly suitable as a domain-specific information system and internet portal

References

Language Technology

- State of the Art and Survey
  - Christian Fluhr:
    - http://www.lt-world.org/HLT_Survey/ltw-chapter8-5.pdf
  - Feiyu Xu
    - http://www.dfki.de/~feiyu/KBIRAF.pdf
  - Doug Oard's Research Page
    - http://www.glue.umd.edu/~oard/research.html

- Resources
  - http://www.ee.umd.edu/medlab/mlir/mlir.html