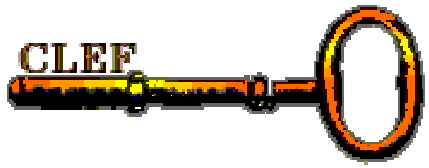
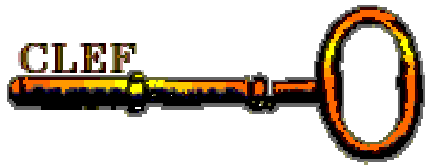


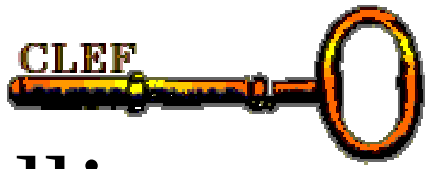
Tools for Multilingual Information Access

Martin Braschler
Eurospider Information Technology AG
8006 Zürich, Switzerland
braschler@eurospider.ch



Types of Tools

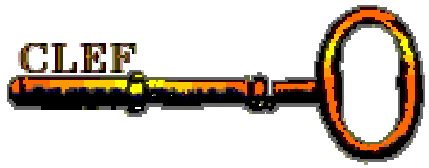
- Mark-Up Tools
- Character Set/ Font Handling
- Language Identification
- Word Segmentation
- Stemming/ Normalization
- Phrase/ Compound Handling
- Entity Recognition
- Terminology Extraction
- Part-Of-Speech taggers
- Parsers/ Linguistic Processors
- Indexing Tools
- Lexicon Acquisition
- Text Alignment
- MT Systems
- Speech Recognition/OCR
- Summarization
- Visualization



Character Set/Font Handling

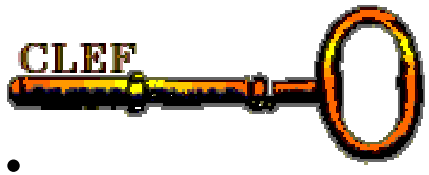
- Input and Display Support
  - Special input modules for e.g. Asian languages
  - Out-of-the-box support much improved thanks to modern web browsers
- Character Set/File Format
  - Unicode/UTF-8
  - XML

Eurospider



Language Identification

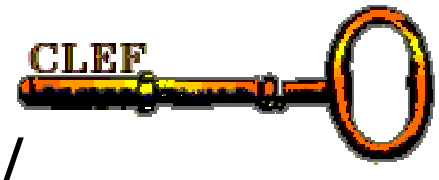
- Different levels of multilingual data:
  - In different subcollections
  - Within subcollections
  - Within items
- Different approaches
  - Trigram
  - Stopwords
  - Linguistic analysis



Stemming/Normalization

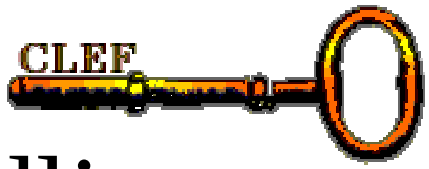
- Reduction of words to their root form
- Important for languages with rich morphology
- Rule-based or dictionary-based
- Case normalization
- Handling of diacritics (French, ..)
- Vowel (re-)substitution (e.g. semitic languages, ..)

Eurospider



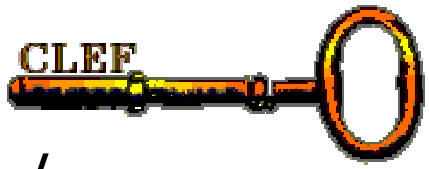
Entity Recognition/ Terminology Extraction

- Proper Names, Locations, ...
  - Critical, since often missing from dictionaries
  - Special problems in languages such as Chinese
- Domain-specific vocabulary, technical terms
  - Critical for effectiveness and accuracy



Phrase/Compound Handling

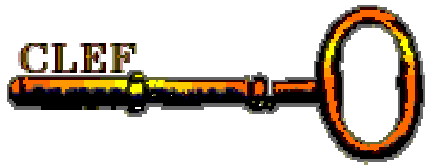
- Collocations („Hong Kong“)
  - Important for dictionary lookup
  - Improves retrieval accuracy
- Compounds („Bankangestelltenlohn“ – bank employee salary)
  - Big problem in German
  - Infinite number of compounds – dictionary is no viable solution



Lexicon Acquisition/ Text Alignment

- Goal: automatic construction of data structures such as dictionaries and thesauri
  - Work on parallel and comparable corpora
  - Terminology extraction
  - Similarity thesauri
- Prerequisite: training data, usually aligned
  - Document, sentence, word level alignment

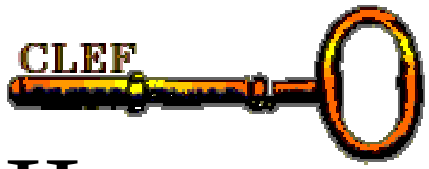
Eurospider



Problem/Challenges

- Availability of Tools
  - Need for good repositories
  - Much work still needed for „exotic“ languages
- Price
  - Some public domain tools
  - Tools can be extremely expensive

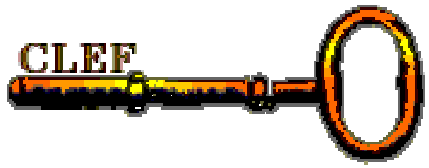
Eurospider



Problem/Challenges II

- Interoperability
  - Too many tools are „black boxes“
  - Need for input/output standards
- Maintenance
- Evaluation
  - Effects of individual components on the system as a whole
  - Test procedures/collections

Eurospider



Tools and CLEF

- Newcomers need a basic tool set to participate
  - Stemmers, language identifiers, handling of diacritical characters, ...
- CLEF wants to be a forum that helps understand the effect of individual tools
  - Input from participants needed wrt used tools/resources
  - Careful analysis of sub-aspects of systems wrt tools is particularly welcome
- CLEF intends to build a tool repository

Eurospider