Multilingual discovery in Europeana

Sjoerd Siebinga, Europeana Office

TrebleClef knowledge transfer day, Berlin 08-12-2009
About this presentation
• Training your system requires high-quality annotated corpora
• Dialectal and cultural differences are found on every level
• European Bureau for Lesser Used Languages (EBLUL)
5 levels of multilingual support

- Multilingual interface
- Multilingual search on a monolingual baseline
- Multilingual browsing of vocabularies
- Multilingual query translation
- Multilingual result translation
Is multilingual support still important?

• Know your users.
• Most web users are mainly monolingual
• Every website offers a service
Scope, Expectations, and Granularity
What can we learn from Europeana
What is a sustainable internet business model

• Do not drink from the same source as your primary customers
• Every web-page view costs money, so a conversion to income of a percentage is key to survival and scalability
• Can you afford to succeed.
• What does it cost to extend coverage for:
  • functionality
  • languages
Metadata is really challenging to work with
What are the basics you need

- bilingual dictionaries
- stemmers
- syntactic parsers
- Named Entity Recognition software
- stop-word list
- language and boundaries detection
- annotated lemmata for the most frequent words
How would you ideally go about your work

- identify the language and/or boundaries
- apply stemmers and syntax parsers
- do Named Entity recognition
- determine dictionary coverage of unique types
- semantic reification
- etc.
What data does Europeana get?
Implications for search engine

- one single index
- no stop-word lists
- no stemming
- no language specific processing pipelines
How should you maintain it

- Have quarterly quality checks with user groups
- Hire experts to update and maintain your bilingual dictionaries
- Quality Assurance is an inherently iterative process
The final frontier: Time, and spelling changes

• Spelling and language evolution exponentially increase the complexity of resources to maintain.

• Which languages are you going to support

• Europeana wants to offer Transcendent Discovery through space and time.
  • Graceful discovery and recovery.
    • Names, words, places, boundaries, countries, etc. Every thing changes over time. We want to make this available through an intuitive interface.
    • Event based browsing is key
MLIA is not difficult, just very expensive

Not development cost, but maintenance cost are the real bottleneck

MLIA is all about resources and a little about technology

Be extremely explicit in what you are going to offer and be certain that the User really wants it, i.e. user-needs should drive business requirements

User Interface and Interaction design is more important than the number of features you offer