Co-funded by the European Community Programme eContent*plus*

europeana connect

Europeana Language Resources Repository

Europeana v1.0 WP3 Meeting

Amsterdam, 14 October 2010

Vivien Petras (HUB)

Coordinated by the Austrian National Library Österreichische Nationalbibliothek

Make Europeana talk European

- Necessary Resources = EuropeanaConnect WP2 tasks (mapping vocabs / query translation)

  → Listing of components & suggested resources:

  - Stop word lists
  - Linguistic analyzers (tokenization, lemmatization, decompounding, multi-word detection, part of speech tagging)
  - Named entity recognizers
  - Translation dictionaries
  - Language identifier

French	English	Spanish
German	Italian	Polish

Dutch	Portugese
Hungarian	Swedish

Functions of the Language Resource Repository

→ infrastructure to support ingesting, storing, maintaining, and providing access to linguistic resources

- Store data in repository with associated code, documentation
- Write code to expose data via common API
- Make corrections, additions
- Integrate new versions
- Deploy resources into production

APIs

- Stop words: no API, just simple text files
- LanguageGuesser
- TranslationDictionary
- LinguisticAnalyzer
  - tokenization, lemmatization, decompounding, multi-word detection, part of speech tagging, named entity recognition

Repository of Language Resources

- Online in EuropeanaLabs
- Java APIs to access each type of resource (wrappers)
- Standardized set of description features for each resource
- Some test corpora for evaluation
- Open-source (anon. download) & password-protected sections (proprietary resources)
- Stop word lists / language identifier / morphological analyzers for all 10 lang.; NE recognizers for en, fr, de; translation dictionaries for 7 lang.

Resources Register

- Publicly-visible register of available resources (http://europeanalabs.eu/wiki/LinguisticResourceRegister)

Language Resource Repository - Morphological Analyzers

1. TreeTagger English
   The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It combines language independent Executable code (without source) and language specific parameter files.
   - Path : https://europeanalabs.eu/core/svn/contrib/lrr/trunk/morphological-analyzers/treetagger/en
   - Type :
     - morphological analyzer
   - License : proprietary
   - Origin : Institute for Computational Linguistics of the University of Stuttgart (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/)
   - Licence notes :
     Free for evaluation, research, and teaching
   - Contact :
     - Helmut Schmid (Helmut.Schmid@ims.uni-stuttgart.de)
   - Fee : (negotiations pending)
   - API : yes
   - Status : adapted
   - Language :
     - English (en)
   - Function :
     - lemmatization
     - POS tagging
   - Accuracy : 96.81% (reported in http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf)
   - Platform :
     - Linux
     - MacOS
     - Windows
   - Speed : 8,000 tokens per second (reported in http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf)
   - Comment : The tagset is an undocumented extension of the Penn-Treebank tagset. The latter is documented in http://www.ims.uni-stuttgart.de/projekte/corplex /TreeTagger/Penn-Treebank-Tagset.pdf . The adaptation of TreeTagger to the Europeana API currently has the following limitation/bug: the token offsets that are generated are false (start offset is always 0).

Language Identifier

- Online in EuropeanaLabs (also as web service: http://www.cross-library.com/FAUSS/)
- Combination metric (Corpus Based, Character Model Based, Function Word Based, Prior Guess Based)
- For all 10 languages + Greek

→ Available from Europeana Language Resources Repository

Wiki: Suggest more Resources

- Open-source language resources: please enter into Wiki: http://europeanalabs.eu/wiki/WP2LanguageResources

Help Europeana to make its Content accessible in different Languages

We are looking for **Open Source Language Resources and Tools**, which can support the efforts in EuropeanaConnect to translate queries into 10 European languages:

Core Languages:	English	French	German	Italian	Polish	Spanish
Secondary Languages:	Dutch	Hungarian	Portugese	Swedisch		

Query translation and metadata translation depend on language resources, which are data or tools for natural language processing, i.e. to detect a language of query, recognize named entities or translate queries. We are looking for:

- dictionaries
- written and spoken corpora
- controlled vocabularies or ontologies
- lexicons, terminologies or other term lists
- tools for language processing, i.e. language detectors, stemmers, lemmatizers, named-entity-detectors, POS-Taggers

Please add any open source resources you might know or work with to the table. We will evaluate them to assess their potential and incorporate them into the Europeana Language Resource Repository.

Name	**Type of Resource**	**Languages**	**Contact Information**	**URL** (if available)
Hungarian Wordnet	Lexicon / Knowledge Source	Hungarian	University of Szeged, Department of Informatics, Human Language Technology Group	http://www.inf.u-szeged.hu/projectdirs/hlt/index_en.html