GROBID–DICTIONARIES

Mohamed Khemakhem
MOHAMED.KHEMAKHEM@INRIA.FR

ELEXIS VIENNA, FEBRUARY 2019
Context

- Started in June 2016
- Intersection:
  - PARTHENOS, Nénufar, BASNUM, DISCO, Paris-Time Machine, DISCO, ELEXIS..
  - TEI-Lex 0, ISO-24613 (LMF)
- PhD Thesis
  - Standard-based Lexical Models for Automatically Structured Dictionaries
Need: Scale-up Parsing Digital Dictionaries

- Digitisation
  - Digitally Born
  - Digitised

- Content
  - Lexical
  - Encyclopedic

- Time
  - Modern
  - Legacy (retro-digitised)
Dictionary Parsing: Landscape

- Domination of rule based methods
  - Ad-hoc
  - Hard/Impossible to scale up

- Just 2 previous works used machine learning
  - Crist, S. (2011)
GeneRation Of Blibliographic Data = GROBID
GROBID: GeneRation Of Bibliographic Data

- Java library for parsing and extracting bibliographic information (mainly PDF)
- Cascading supervised machine learning
  - Sequence Labelling: Conditional Random Fields (CRF)
- TEI-compliant output
- Users: Research Gate, HAL, Mendeley, CERN,...
GROBID-Family

GeneRation Of Bibliographic Data = GROBID

GROBID-Dictionaries

GROBID-NER (Entity Fishing)

GROBID-Quantities

GROBID-Astro
GROBID-Dictionaries: Approach
Dictionary Segmentation Level
condenser [kəˈdɛsə] v. t. (lat. condensare, rendre épais). Rendre plus dense, réduire à un moindre volume. | Liquéfier un gaz par refroidissement ou compression : le froid condense la vapeur d'eau. | Fig. Exprimer d'une manière concise, en peu de mots :
<table>
<thead>
<tr>
<th>Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF model</td>
</tr>
<tr>
<td>Segmented Page</td>
</tr>
<tr>
<td>CRF model</td>
</tr>
<tr>
<td>Segmented Page</td>
</tr>
<tr>
<td>CRF model</td>
</tr>
<tr>
<td>Segmented LE</td>
</tr>
<tr>
<td>CRF models</td>
</tr>
<tr>
<td>Segmented Form/Sense</td>
</tr>
<tr>
<td>CRF model</td>
</tr>
</tbody>
</table>
Output Serialisation

- TEI-Lex-0
  - New recommendations for unified TEI modelling

- LMF
  - Reviewed meta model
  - A new TEI-based serialisation: ISO 24613-4
Activation: MATTER approach
Activation: MATTER approach
Enhanced Usability: Setup+Annotation

- Khemakhem et al. 2018a
Samples: Lexical Dictionaries

- Khemakhem et al. 2017
  - Monolingual digital English dictionary
  - Bilingual digitised Fang-French dictionary
Samples: Encyclopedic Dictionaries and MAC

- Khemakhem et al. 2018b

Source: Manuscript Auction Catalogues (1889)
Work in Progress

- Feature engineering
  - Improve for existing models
  - Models for a category of dictionaries
  - Varied Data is crucial
Work in Progress

- Feature engineering
  - Improve for existing models
  - Models for a category of dictionaries
  - Varied Data is crucial

- Implement new models
  - `<sense>`, `<xr>`, `<re>`...
  - Input from TEI-Lex 0 WG
Work in Progress

- Feature engineering
  - Improve for existing models
  - Models for a category of dictionaries
  - Varied Data is crucial
- Implement new models
  - `<sense>`, `<xr>`, `<re>`..
  - Input from TEI-Lex 0 WG
- Etymology extension
References

- **Sean Crist (2011)**. Processing the text of bilingual print dictionaries
- **Mohamed Khemakhem, Axel Herold and Laurent Romary (2018a)**. Enhancing Usability for Automatically Structuring Digitised Dictionaries. GLOBALEX workshop at LREC 2018, May 2018, Miyazaki, Japan. 2018
- **Mohamed Khemakhem, Laurent Romary, Simon Gabay, Hervé Bohbot, Francesca Frontini, et al. (2018b)**. Automatically Encoding Encyclopedic-like Resources in TEI. The annual TEI Conference and Members Meeting, Sep 2018, Tokyo, Japan
- **Mohamed Khemakhem, Carmen Brando, Laurent Romary, Frédérique Mélanie-Becquet and Jean-Luc Pinol (2018c)**. Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories. JADH2018 "Leveraging Open Data", Sep 2018, Tokyo, Japan
- **David Lindemann, Mohamed Khemakhem and Laurent Romary (2018)**. Retro-digitizing and Automatically Structuring a Large Bibliography Collection. European Association for Digital Humanities (EADH) Conference, December 2018, Galway, Ireland
thank-you
[thangk-yoo]

verb (used with object)
1. to express gratitude, appreciation, or acknowledgment to:
   She thanked them for their hospitality.

noun
2. Usually, thanks, a grateful feeling or acknowledgment of a
   benefit, favor, or the like, expressed by words or otherwise:
   to return a borrowed book with thanks.