LEX 1 & LEX 3

Simon Krek (Jožef Stefan Institute, Ljubljana, Slovenia)
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
LEX 1 – segmentation, conversion tools

• The LEX1 infrastructure will be set up during the project and will consist of several sets of tools:

• (1) tools for automatic segmentation and structuring of lexicographic content for dictionaries that are currently produced in digital environments and are encoded in their own custom data format.

• (2) conversion and alignment tools that will provide users of the infrastructure with the possibility to harmonise and convert their lexicographic resources to a uniform data format that allows their seamless integration in Linked Open Data.
LEX 1 – linking tools

• (3) linking tools that will provide conceptual interoperability, enabling linking of ELEXIS lexicographic resources.

• This will provide the possibility to link lexical entries, senses and fundamental concepts in different lexical resources, using a semi-automatic approach.

• BabelNet, as an existing multilingual resource to provide cross-lingual linking, will be exploited for this purpose.
LEX 1 – matrix dictionary = dictionary matrix

• (4) **ELEXIS matrix dictionary**: extensive linking of existing lexicographic resources by pivoting through BabelNet will enable the creation of what we call ELEXIS matrix dictionary –

• a universal **repository of linked senses**, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in all types of existing lexicographic resources, monolingual, multilingual, modern, historical etc., available through a RESTful web service as **part of LEX1**.

• Freshly renamed: **DICTIONARY MATRIX – a matrix of dictionaries**
LEX 3 infrastructure

- LEX3 infrastructure will be dedicated to retrodigitised dictionaries and will include

- (1) tools for automatic segmentation and structuring of content in retro-digitised dictionaries, and

- (2) Legacy Dictionary Viewer: a generic, modular dictionary publication tool for retrodigitised dictionaries which offers interfaces for the analysis and profiling of the underlying lexical data.
  - The viewer will allow users — lexicographers and digital humanists alike — to publish user-friendly, online editions or retrodigitised dictionaries without facing extensive technical or financial obstacles.
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
GROBID-dictionaries

• GROBID-Dictionaries is a GROBID sub-module, implementing a java machine learning library, for structuring digitized lexical resources and entry-based documents with encyclopedic or bibliographic content.

• It allows the parsing, extraction and structuring of text information in such resources.

• Presentation: Mohamed Khemakhem
• On GitHub: https://github.com/MedKhem/grobid-dictionaries
• ELEXIS installation: http://grobid.elex.is/
Conversion and alignment

ELEXIS (XML schema) conversion tools

- SSSJ
- ANW
- DDO

→ TEI-Lex0

GROBID-dictionaries

- Word
- PDF
- OCR

→ TEI-Lex0

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
Conversion and alignment – TEI-Lex0

Basic conversion (quick & easy) 

Full conversion (slow & thorough)
Conversion tools - process

• Quick and easy
  • Upload your dictionary in custom XML format -> parse (or provide schema)
  • **Identify basic dictionary elements defined by ELEXIS data model**
  • Convert your dictionary to TEI-Lex0 and download

• Slow and thorough
  • Upload your dictionary in custom XML format -> parse (or provide schema)
  • **Align your schema and TEI-Lex0**
  • Convert your dictionary to TEI-Lex0 and download
Conversion and alignment tools

• **(E)LEXIFIER**
  • *Lexifier* - the dominant language of a pidgin or creole that serves as the basis for most of its vocabulary (Wiktionary)
  • The tool identifies only basic dictionary elements (entry, headword, sense, definition, example etc.) and produces a file in valid TEI-Lex0 format with original dictionary content.
  • XSLT tools for “complete” conversion
ELEXIS DICTIONARY MATRIX

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
BabelNet – five concepts FANCY-n

**Noun**

- **fancy, fantasy, illusion**
  - Something many people believe that is false
  - ID: 00033016a | Concept

- **fancy**
  - A kind of imagination that was held by Coleridge to be more casual and superficial than true imagination
  - ID: 00033017a | Concept

- **fancy, fondness, partiality**
  - A predisposition to like something
  - ID: 00033018a | Concept
Repository of senses – “matrix dictionary”

CONCEPT 1 - “FANCY-n”: something many people believe is false

CONCEPT 2 “FANCY-n” CONCEPT 2: a predisposition to like something
Repository of senses - SloWNet

CONCEPT 1 - “FANCY-n”: something many people believe is false

CONCEPT 2 “FANCY-n” CONCEPT 2: a predisposition to like something
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
Presentations

• GROBID-Dictionaries

• ELEXIS data model

• Linking & dictionary matrix

• Word Sense Disambiguation

• Mohamed Khemakhem
  • Inria-ALMAnaCH

• Carole Tiberius
  • Dutch Language Institute

• John McCrae
  • National University of Ireland

• Roberto Navigli
  • Sapienza University of Rome
Timeline – LEX1 / 3

• **T1.4** Data Structuring for existing resources: **M1-M24**
  - GROBID-dictionaries, ELEXIFIER
    - **D1.3** Tools for the automatic segmentation and identification of lexicographic content (**M24**) (software)
    - **D1.4** ELEXIS conversion tools (**M36**) (software)

• **T2.2** Semi-automatic linking of lexical resources: **12M-48M**
  - just started → end of project
    - **D2.2** Interoperable interface for Lemon and TEI resources (**M24**) (software)
    - **D2.4** Cross-lingual Lexical Resource Linking Web Service (**M42**) (software)