Linking the Bilingual Latin-English Dictionary Lewis & Short to the LiLa Knowledge Base

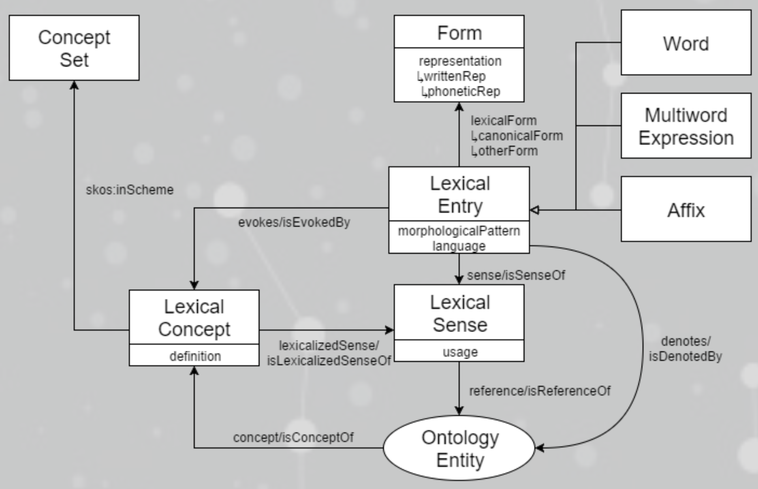
Ginevra Martinelli, Marco Passarotti

CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan

ginevra.martinelli01@icatt.it, marco.passarotti@unicatt.it

LiLa

- The LiLa project builds a Linked Data-based Knowledge Base of Linguistic Resources and Natural Language Processing (NLP) tools for Latin.
- The Knowledge Base consists of different kinds of objects connected via an explicitly-declared vocabulary for knowledge description.
- Interoperability is attained by linking all the entries in lexical resources and tokens in corpora that point to the same lemma to the corresponding lemmas in the Lemma Bank.
- The Lemma Bank is a large collection of Latin lemmas, currently comprising of almost 200,000 canonical forms (approx. 130,000 lexical items).



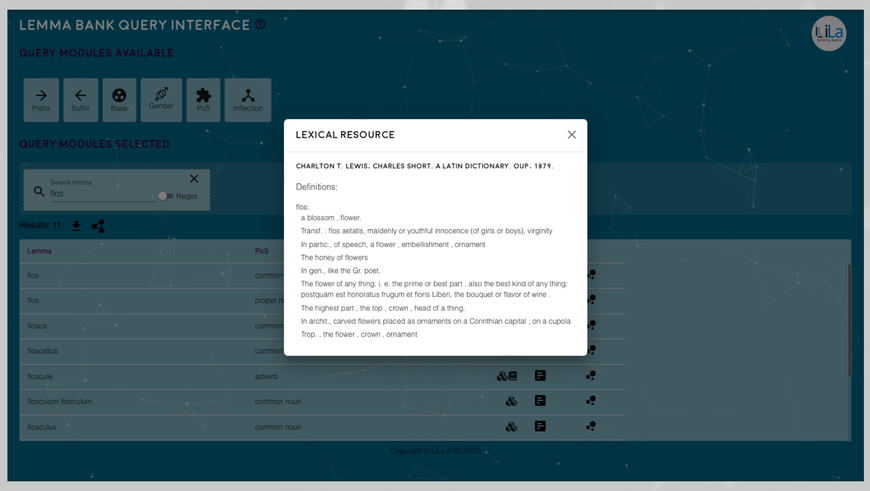
Linking the Lewis & Short

The Dictionary

- A bilingual Latin-English Dictionary, curated by Ch. T. Lewis and Ch. Short (1879).
- Encoded in TEI XML by the Perseus Project and available on a series of desktop and web applications.

Linking the entries

1. Spelling normalisation.
2. Mapping of part-of-speech and inflectional information.
3. Matching on the basis of the tuple written representation - PoS.
4. Results: 31,142 1:1 matches out of the 51,596 total entries of L&S.



1:0 Matches

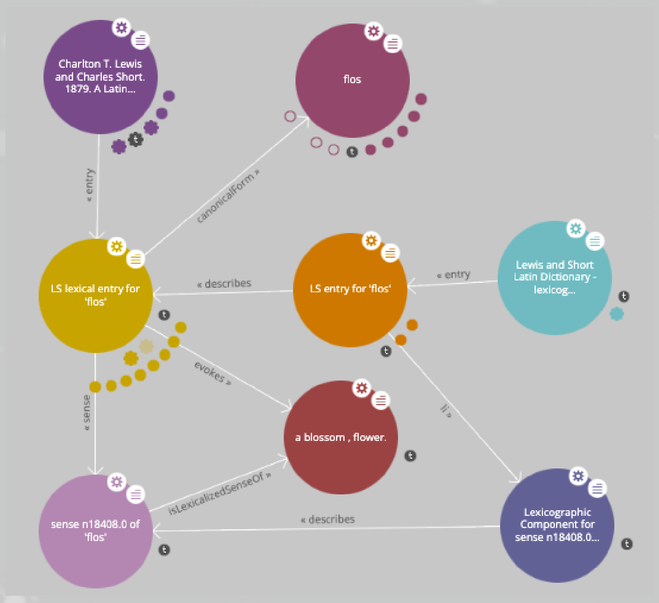
- Entries without a corresponding canonical form in LiLa.
- Affixal forms, graphic indications or crasis.
- Multiword expressions.

- Inflected forms.
- Alternative spellings.
- Entries differing from other entries only in the paradigmatic slot representing the lemma.
- False reads.

```
▼<TEI.2>
 ▼<text>
  ▼<body>
   ▼<div0 type="alphabetic letter" n="L" org="uniform" sample="complete">
    ▼<entryFree id="n27038" type="main" key="Luca bos" opt="n">
      <orth extent="full" lang="la" opt="n">Lūca bos</orth>
      , v. Lucani, D.
     </entryFree>
    </div0>
   </body>
  </text>
</TEI.2>
```

Modelling

- The "OntoLex lexicography module" or `lexicog` is used to capture the structural information expressed in a lexicographic resource.
- Lexical, or linguistic information are represented by the classes and properties of Ontolex-Lemon.
- `Lexical entry` is an item in the lexicon of a given language.
- `Lexicographic entry` is a record in a linguistic resource that documents or discusses some properties of a given lexical item.
- Lexicographic entries are a special subset of a larger class called `Lexicographic Component`; components can be used to represent senses, sense groups or subentries.
- The property `lexicog:describes` provides a link between lexical, linguistic and structural information.



Useful Links

- LiLa's SPARQL endpoint: `https://lila-erc.eu/sparql`
- LiLa's Query Interface: `https://lila-erc.eu/query`
- CIRCSE L&S GitHub: `https://github.com/CIRCSE/LewisShort`
- LiLa's L&S Lexical Resource: `https://lila-erc.eu/data/lexicalResources/LewisShort/Lexicon`

https://lila-erc.eu

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, Grant Agreement No. 769994