Converting MWE lexicons into LMF

Tristan Mollet

Internship Feb-May 2017

Supervised by Núria Gala and Carlos Ramisch

Adapted and presented by Carlos Ramisch

Lexicon development

- Use of specialized tools and file formats
- Spreadsheets and exported TSV files
  - Tab-separated values in columns
  - Easy to generate and manipulate
  - Hard to share, maintain and structure

Problems of TSV lexicons

- Semantics of each column and value
- Traceability of information
  - Sources (auto, manual), versions
- Redundancy
  - Lack of structure
- Sharing and interoperability

Context

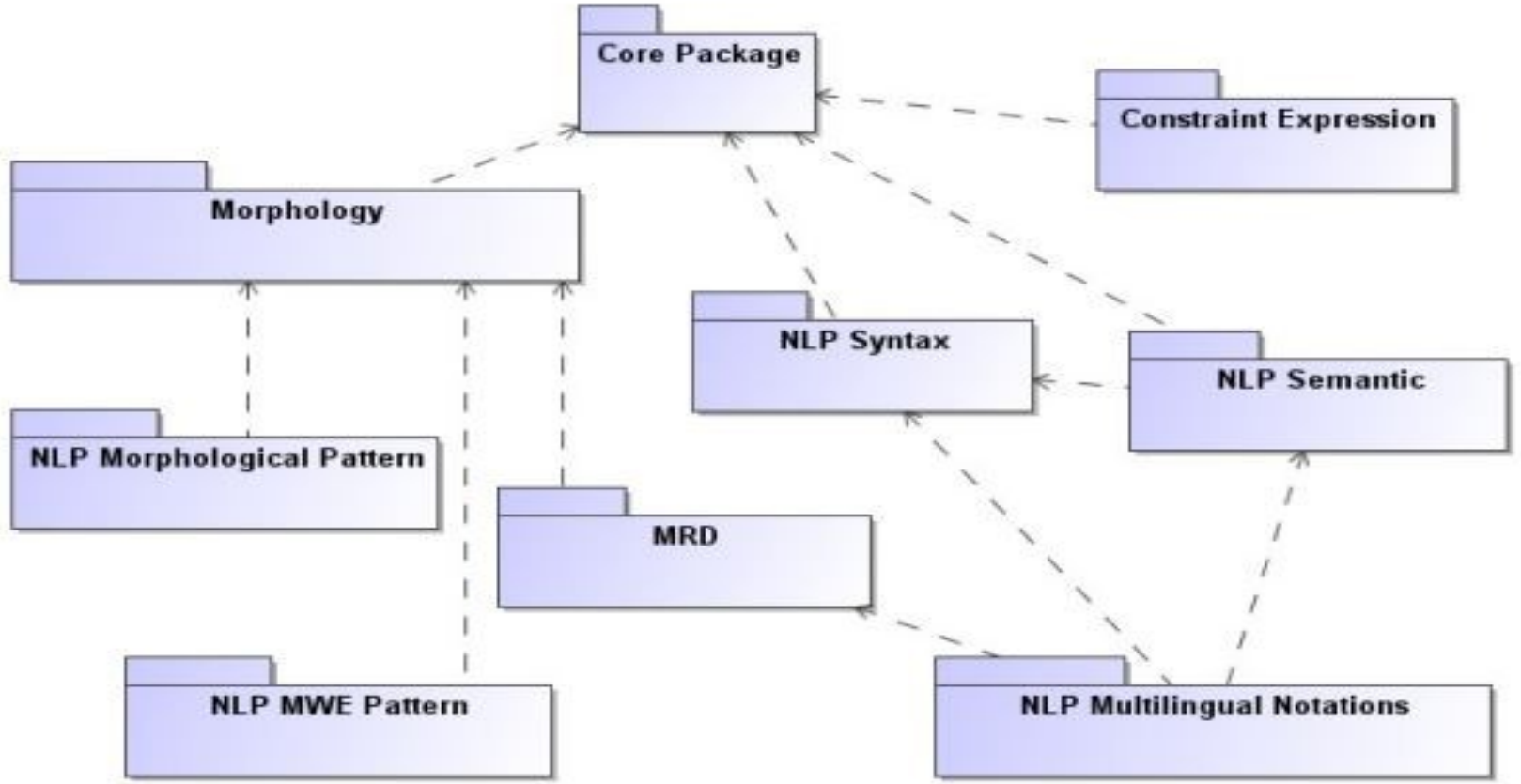
- ReSyf: lexicon of French with lexical units grouped into synsets and graded according to simplicity
- Compositionality datasets: nominal compounds annotated for compositionality degree
- DeQue: lexicon of complex prepositions and conjunctions in French

  - All include MWEs and use TSV + README files

Goals of the internship

- Define a format to solve the limitations of TSV
- Create a web interface to
  - Import existing TSV lexicons
  - Download converted lexicons in standard format
  - Look up imported lexicons (basic look-up)

Format: LMF


Core Package
Constraint Expression
Morphology
NLP Syntax
NLP Semantic
NLP Morphological Pattern
MRD
NLP MWE Pattern
NLP Multilingual Notations

LMF implementation

- XML
  - Validated by DTD or XML Schema
- **RELISH-LMF** and UBY-LMF
  - Uses XML-Schema for validation

Extensions: source

```
<!-- Source element: contains id and timestamp-->
   <define name="SourceElem">
       <zeroOrMore>
           <element name="me:Source">
               <attribute name="id">
               </attribute>
               <attribute name="timestamp">
               </attribute>
               <zeroOrMore>
                   <ref name="relish.lmf.fs"/>
               </zeroOrMore>
           </element>
       </zeroOrMore>
   </define>
```

Extensions: statistics

```
<!-- Statistics element : contains all statistics -->
<define name="StatisticsElem">
    <optional>
        <element name="me:Statistics">
            <zeroOrMore>
                <ref name="relish.lmf.fs"/>
            </zeroOrMore>
        </element>
    </optional>
</define>
```

Example

annotator-id	mwe-id	timestamp	simplest	average	category
alain13090	6	2016-09-15 03:27:29	ressources humaines	gestion du personnel	Personne ou être vivant

```
<Lexicon xml:lang="fr">
       <LexicalEntry xml:id="le1">
           <Lemma type="Form">
               <feat att="simplest" val="ressources humaines"/>
           </Lemma>
           <Sense synset="ss6"/>
       </LexicalEntry>
       <LexicalEntry xml:id="le2">
           <Lemma type="Form">
               <feat att="average" val="gestion du personnel"/>
           </Lemma>
           <Sense synset="ss6"/>
       </LexicalEntry>
       <Synset xml:id="ss6">
           <feat att="category" val="Personne ou être-vivant"/>
           <me:Source id="alain13090" timestamp="2016-09-15T03:27:29"/>
       </Synset>
</Lexicon>
```

Convert TSV → LMF-XML

- Read TSV files
- Transform into Java objects
- Use Java annotations to convert into XML
  - Java API for XML Building – JAX

- **Problem**: matching columns and XML elements

Meta-information file

- JSON file that defines the correspondence
  - LMF element names are the keys
  - TSV column headers are the values
- Converter:
  - Takes TSV + meta-info as input
  - Creates Java objects
  - Generates LMF-XML as output using annotations

```
java -jar TSVtoXMLConverter.jar source.tsv meta-info.json
```

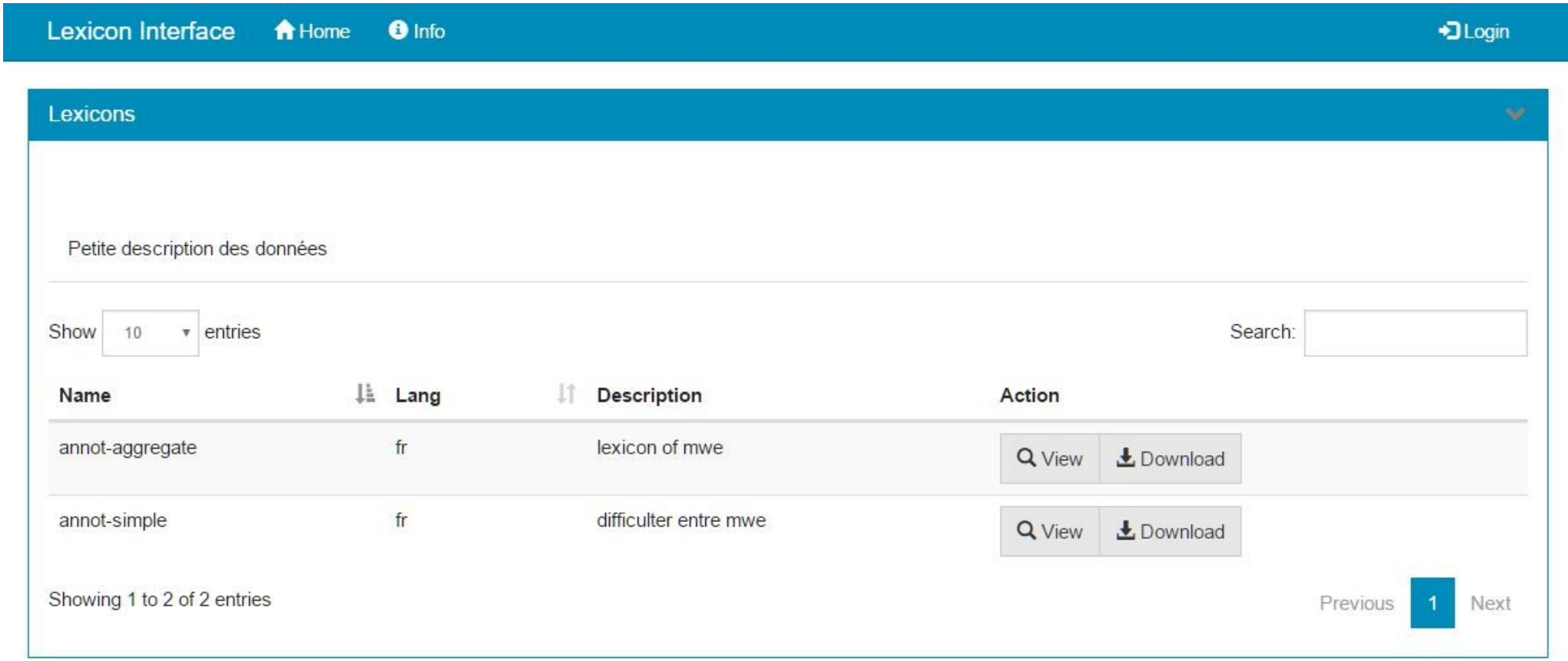
Example: meta-information

```
{
   "LexiconName": "Example",
   "description": "meta-information's file example",
   "Columns": [
       "col1",
       "col2"
   ],
   "Lexicon": {
       "xml:lang": "fr",
       "LexicalEntry": [
           {
               "xml:id": "col1",
               "Lemma": {
                   "feat": [
                       {
                           "att": "exemple",
                           "val": "col2"
                       }
                   ]
               }
           }
       ….
```

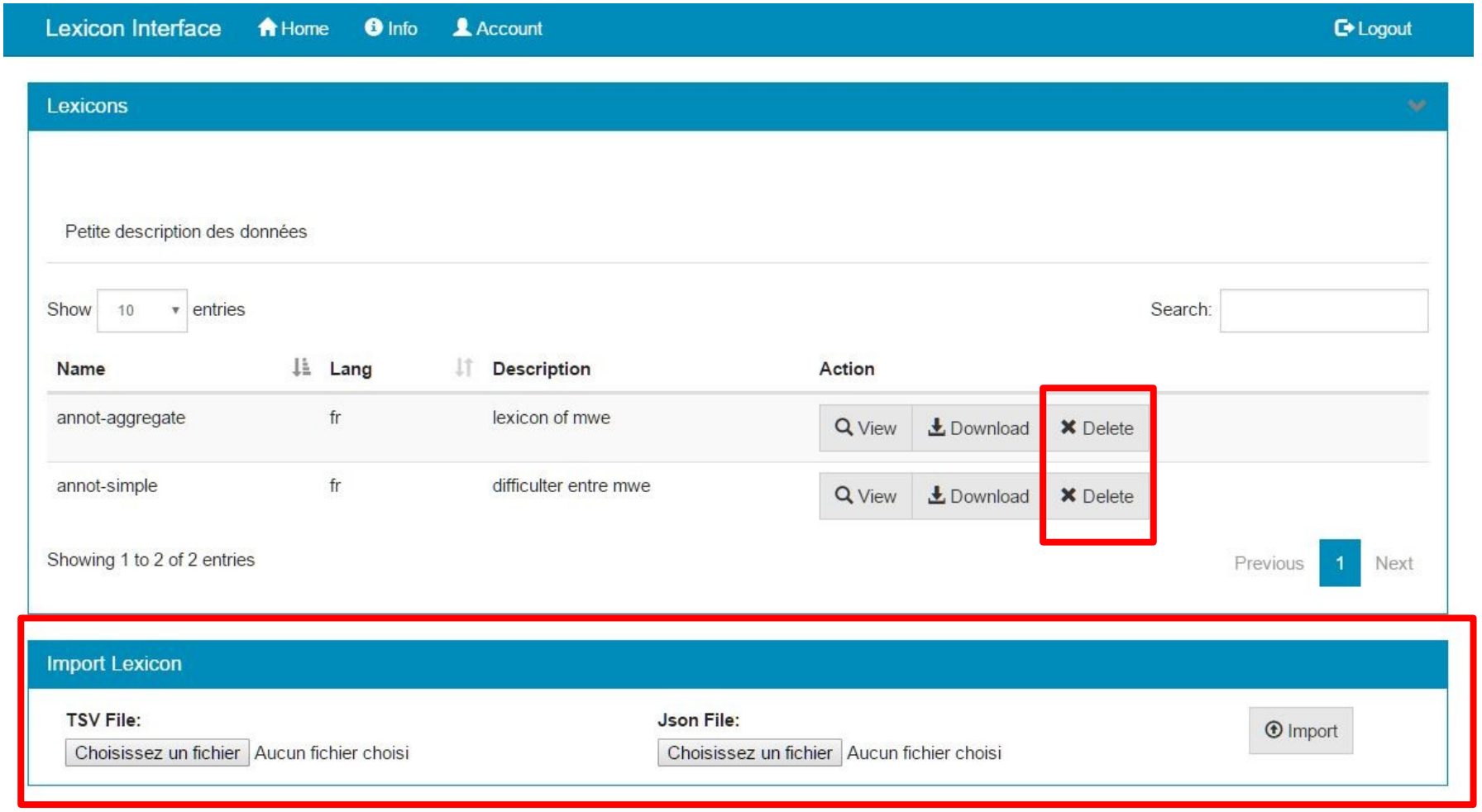
Web interface

- Import a TSV lexicon
  - Load into SQL database
- Export an LMF lexicon
- Look-up an imported lexicon
  - Show entries list
  - Search lemmas
  - Show details of an entry

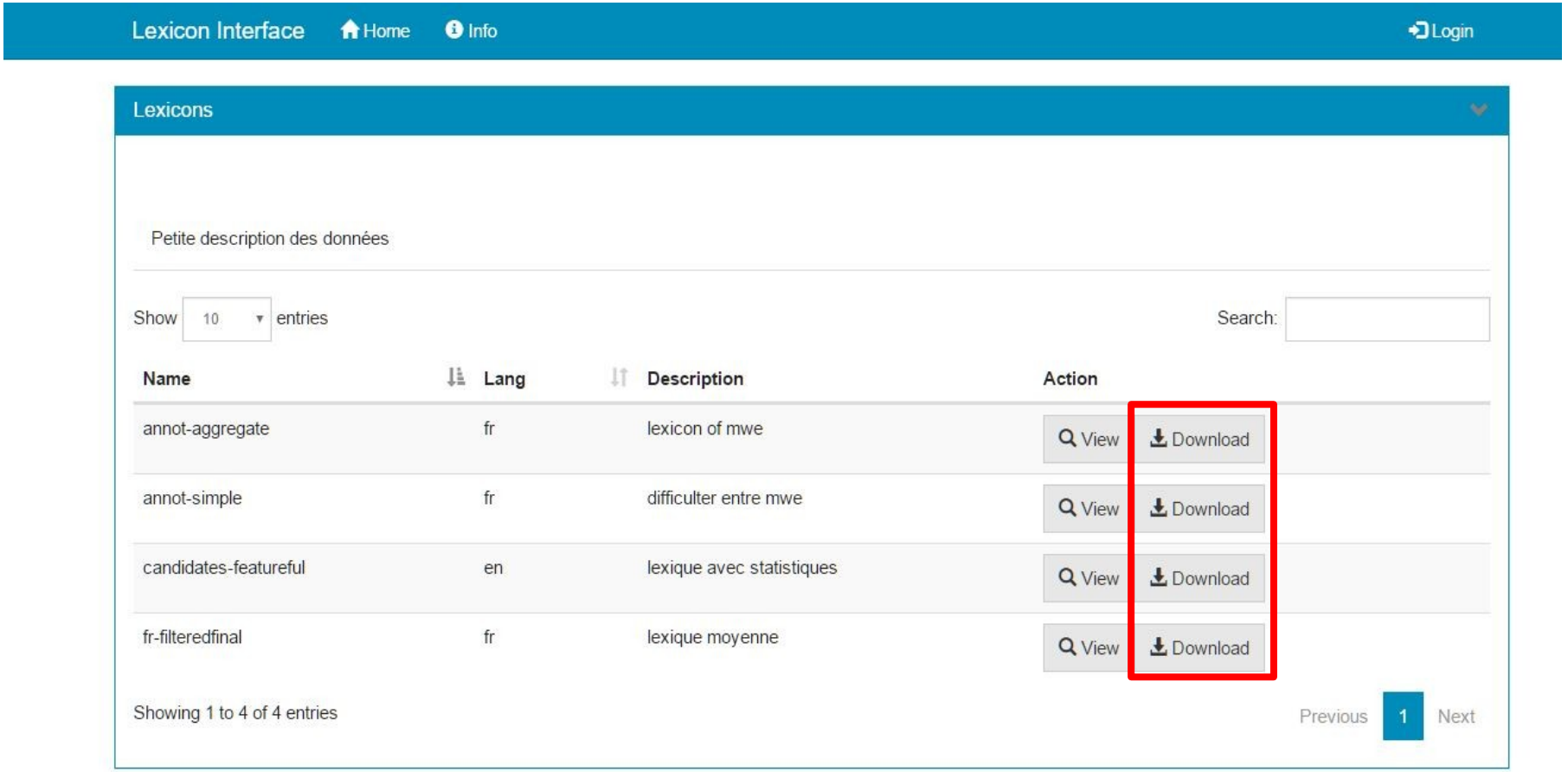
Home page


Lexicon Interface
Home
Info
Login
Lexicons
Petite description des données
Show 10 entries
Search:
Name
Lang
Description
Action
annot-aggregate
fr
lexicon of mwe
View
Download
annot-simple
fr
difficulter entre mwe
View
Download
Showing 1 to 2 of 2 entries
Previous
1
Next

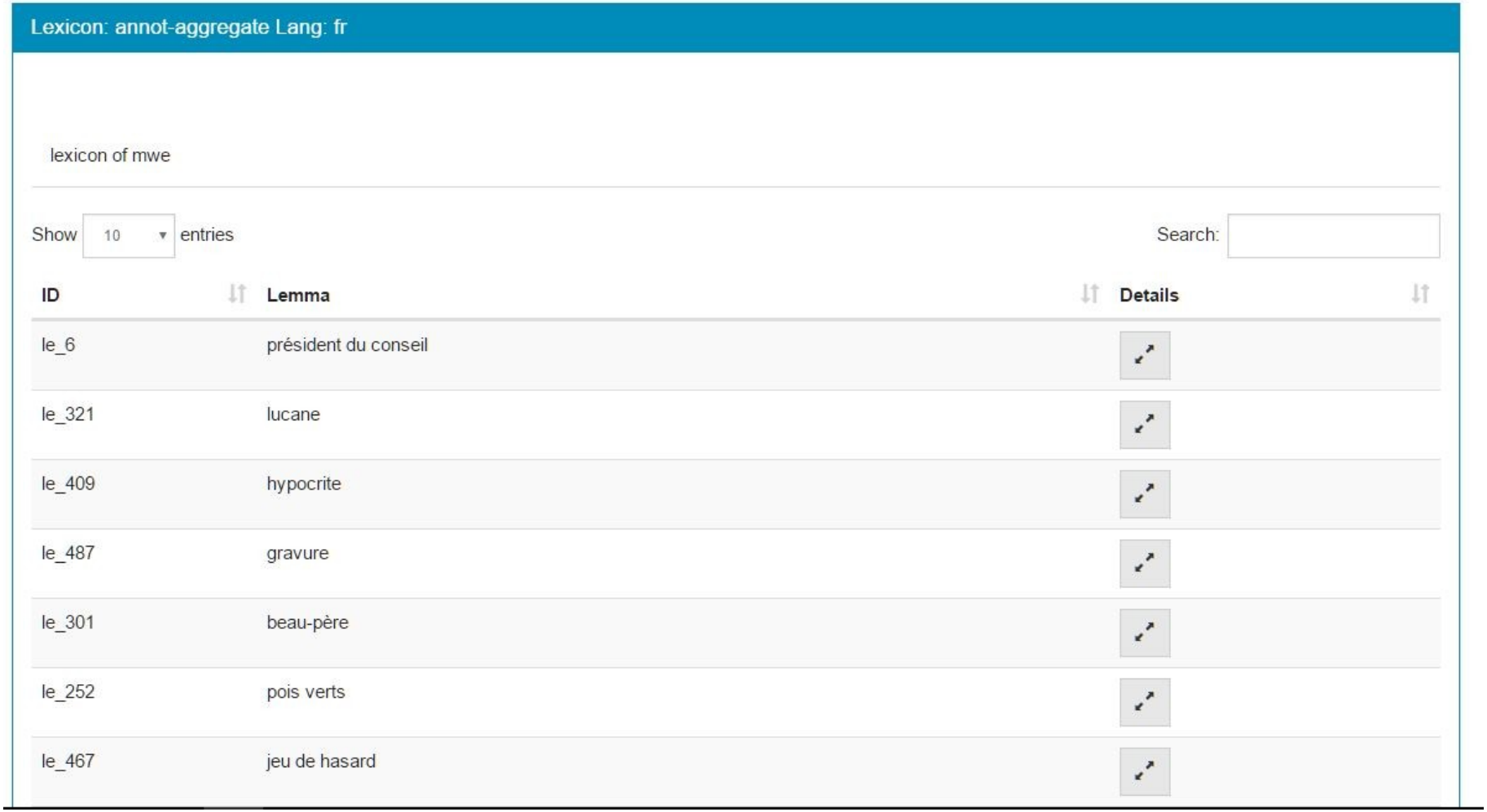
Lexicon import (admin)



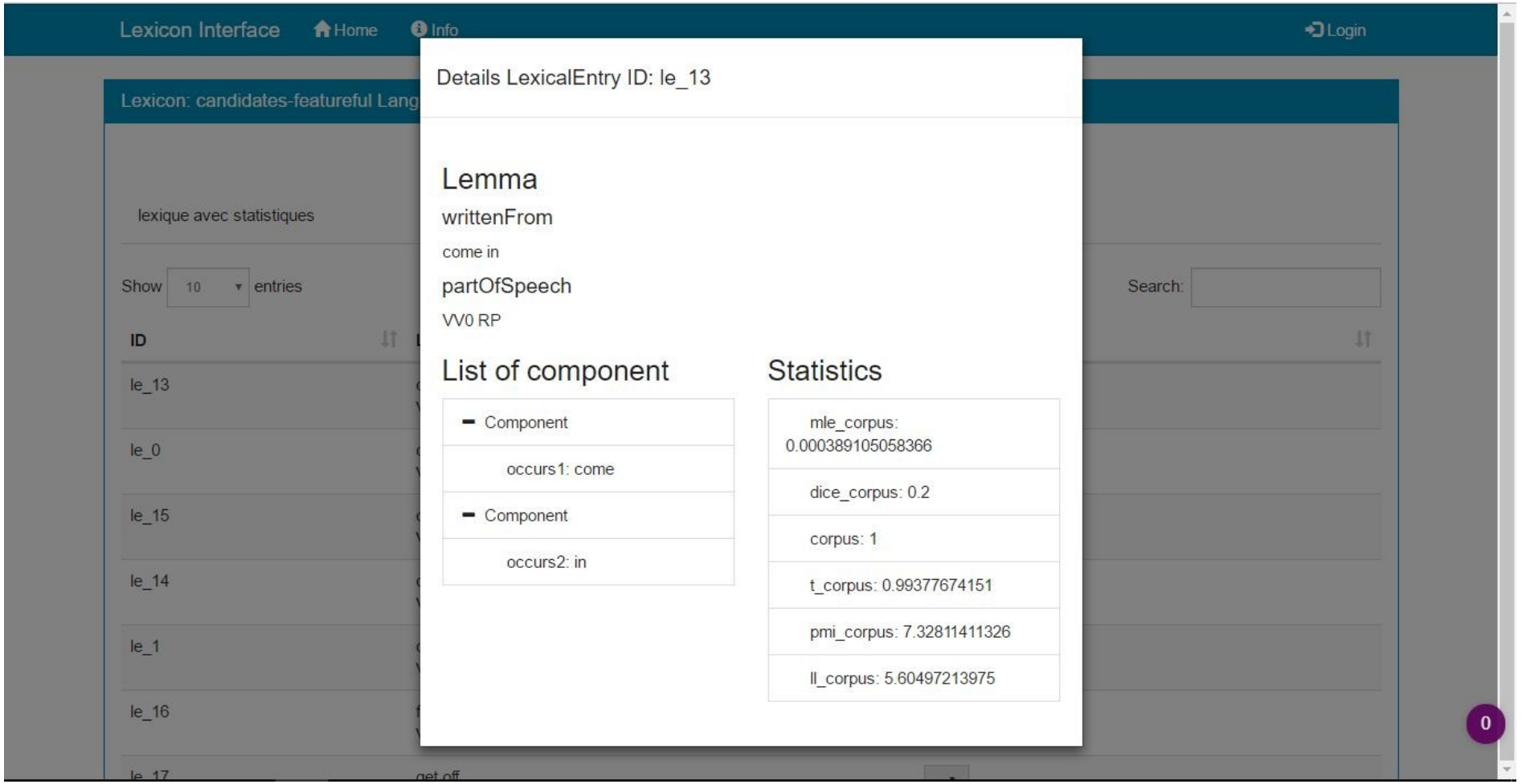
Download LMF lexicon



Lexicon look-up



Entry information



Relevance for PARSEME-FR

- Easy conversion of TSV files
- Minimal look-up interface
- Share PARSEME-FR lexicons (e.g. DeQue)

- Possible evolutions
  - Advanced search
  - Lexicon edition
  - Implement other required LMF elements

https://talep-lexiques.lif.univ-mrs.fr/

Merci !

These slides are based on **Tristan Mollet**'s internship defense. His work described here was carried out at LIF in Feb-May 2017 under the supervision of Núria Gala and Carlos Ramisch