

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

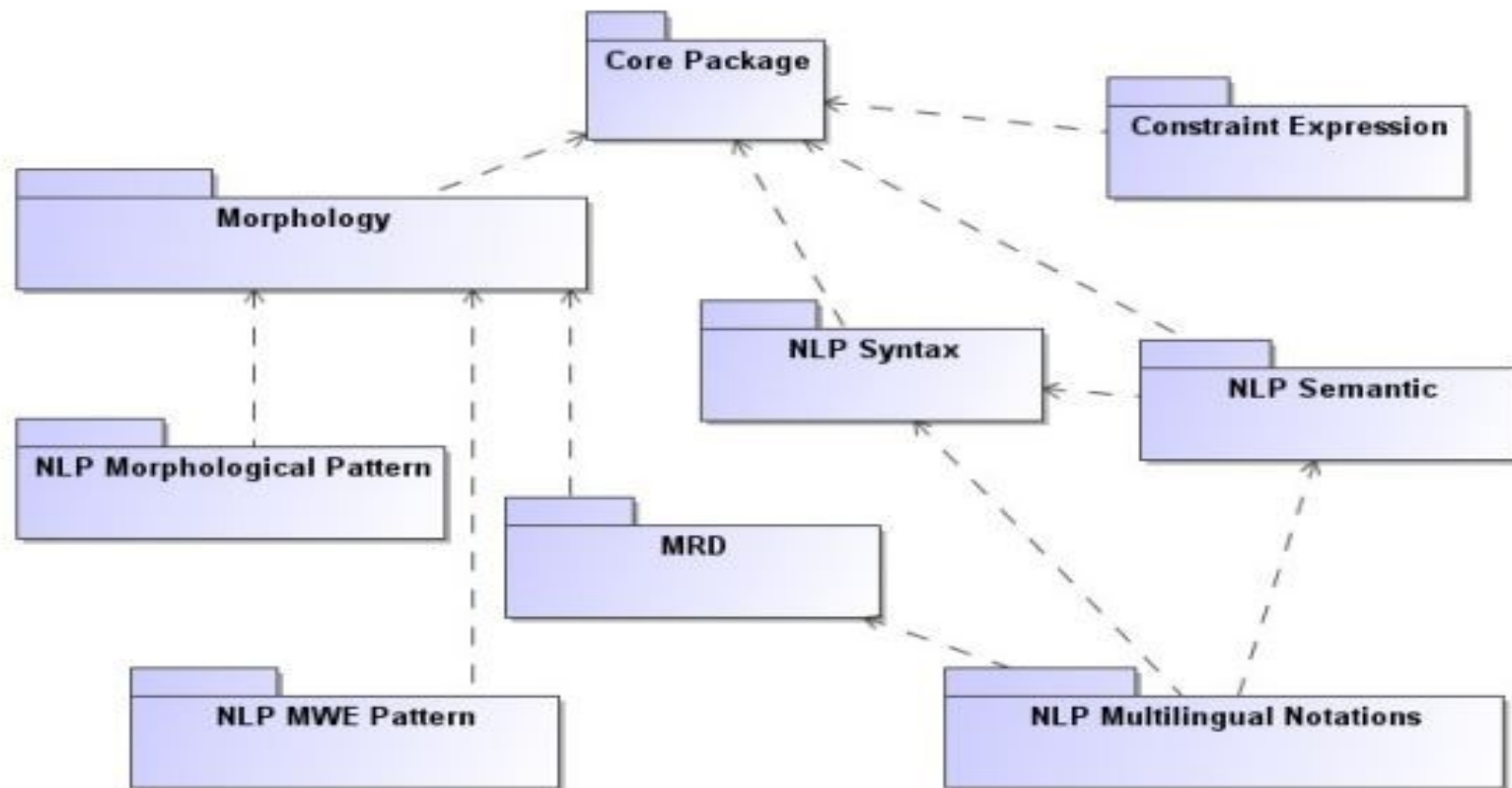
- ReSyf: lexicon of French with lexical units grouped into synsets and graded according to simplicity
- Compositionality datasets: nominal compounds annotated for compositionality degree
- DeQue: lexicon of complex prepositions and conjunctions in French
 - All include MWEs and use TSV + README files



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





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

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

```
<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 – JAXB
- **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

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)

Lexicon Interface [Home](#) [Info](#) [Account](#) [Logout](#)

Lexicons

Petite description des données

Show entries Search:

Name	Lang	Description	Action
annot-aggregate	fr	lexicon of mwe	View Download Delete
annot-simple	fr	difficulter entre mwe	View Download Delete

Showing 1 to 2 of 2 entries Previous **1** Next

Import Lexicon

TSV File: Aucun fichier choisi

Json File: Aucun fichier choisi

[Import](#)

Download LMF lexicon

Lexicon Interface

[Home](#)

[Info](#)

[Login](#)

Lexicons

Petite description des données

Show entries

Search:

Name	Lang	Description	Action	
annot-aggregate	fr	lexicon of mwe	View	Download
annot-simple	fr	difficuler entre mwe	View	Download
candidates-featureful	en	lexique avec statistiques	View	Download
fr-filteredfinal	fr	lexique moyenne	View	Download

Showing 1 to 4 of 4 entries








Previous [1](#) Next

Lexicon look-up

Lexicon: annot-aggregate Lang: fr

lexicon of mwe

Show entries

ID	Lemma	Details
le_6	président du conseil	
le_321	lucane	
le_409	hypocrite	
le_487	gravure	
le_301	beau-père	
le_252	pois verts	
le_467	jeu de hasard	

Entry information

The screenshot displays a web application interface for a lexicon. The main content area is titled "Lexicon: candidates-featureful Lang" and includes a search bar, a "Show 10 entries" dropdown, and a list of entries with IDs like le_13, le_0, le_15, le_14, le_1, le_16, and le_17. A modal window is open over the entry "le_13", titled "Details LexicalEntry ID: le_13". This modal contains the following information:

Lemma
writtenFrom
come in
partOfSpeech
VV0 RP

List of component

- Component
 - occurs1: come
- Component
 - occurs2: in

Statistics

mle_corpus:	0.000389105058366
dice_corpus:	0.2
corpus:	1
t_corpus:	0.99377674151
pmi_corpus:	7.32811411326
ll_corpus:	5.60497213975



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