BFM & TXM

Tools for Corpus Analysis and Publication

Alexei Lavrentiev
IHRIM Research Lab
CNRS
alexei.lavrentev@ens-lyon.fr

DEMM - Digital Edition of Medieval Manuscripts, Lyon

*SLH: Slides by Serge Heiden

16 January 2019

Outline

- BFM corpus and its resources
- TXM Hands-on training
  - Reading, searching and statistics
  - Corpus import
    - XTZ module
  - Data export
    - tables, graphics
    - TEI-TXM format

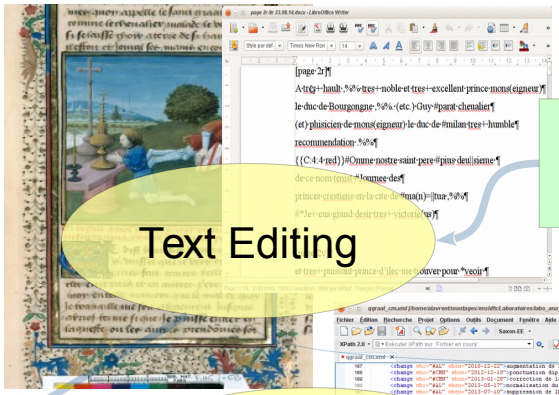
Base de Français Medieval

- http://txm.bfm-corpus.org
- Founded in 1989 by C. Marchello-Nizia
- Corpus of Old and Middle French Texts
  - 9th-15th c.
  - BFM2019 : 5.16 M tokens, 167 texts
  - Verified POS in 39 texts (932 000 tokens)
    - CATTEX 2009 and UD tags
- Most of the texts are digitized critical editions

Base de Français Medieval

- Original digital editions
  - *Queste del saint Graal*
    - ed. C. Marchello-Nizia and A. Lavrentiev
  - *Vie de saint Alexis*
    - ed. T. Rainsford and C. Marchello-Nizia
  - *Image du Monde* (prose) de Gossouin de Metz
    - ed. N. Kanaoka
  - *Quinze joies de mariage*
    - ed. N. Kanaoka

BFM Publishing Workflow



Philological Principles and Encoding Rules

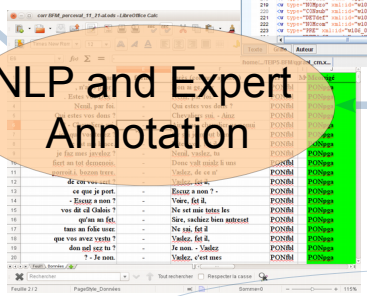
Text Editing

Evaluation (Editorial Board)

Markup and Correction

TEI XML Critical Edition

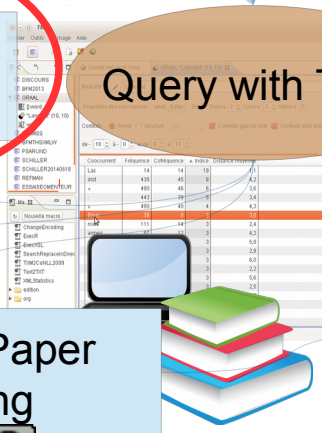
Metadata Database



NLP and Expert Annotation

TEI TXM XML Document

Query with TXM



Virtual Libraries, Corpus collections

Online and Paper Publishing

Research works

BFM Resources

- portal http://txm.bfm-corpus.org
  - online queries, subcorpora, exporting results
  - PDF of the editions
- on request (by email)
  - XML-TEI source files
  - TXM “binary” corpus
    - including TEI-TXM XML files
  - Other formats (TXT, CoNLL-U...)

BFM Resources

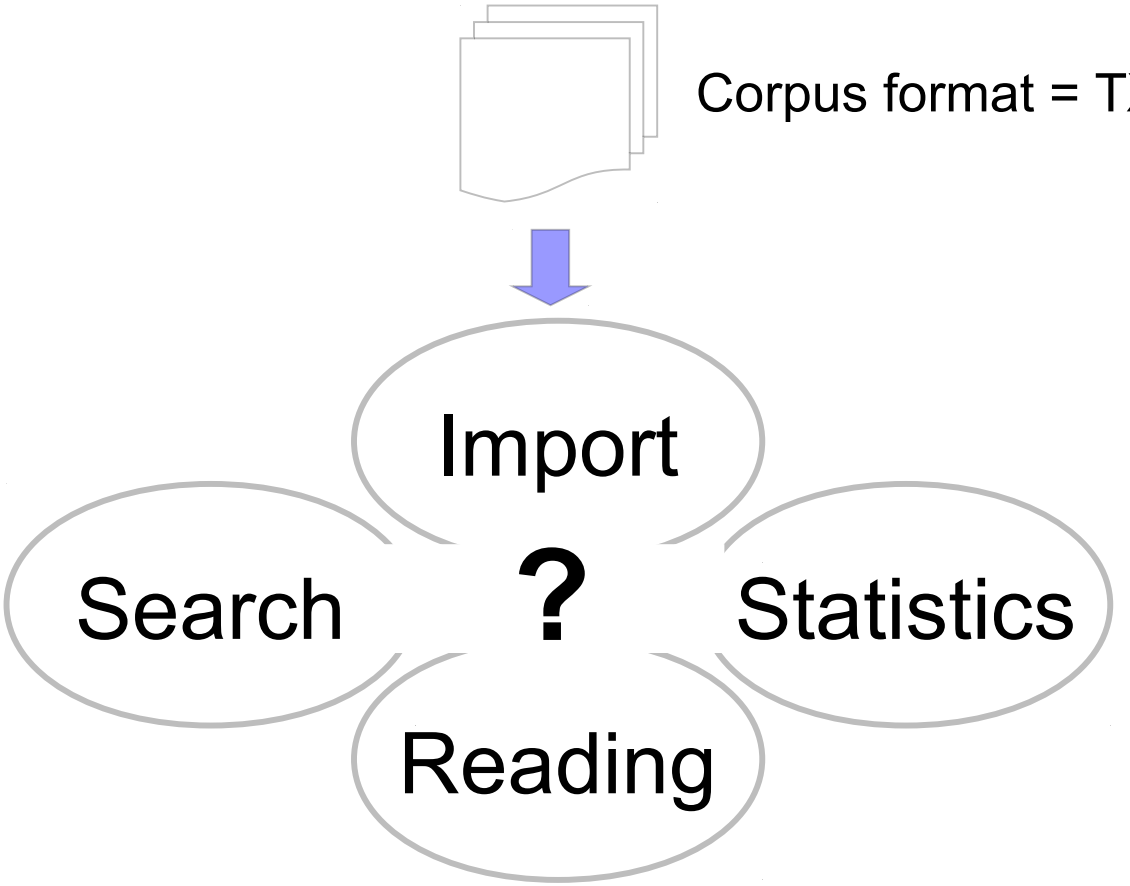
- project website http://bfm.ens-lyon.fr
  - Encoding guidelines, User manuals, etc.
  - FRO.PAR language model for TreeTagger
    - CATTEX 2009 Tagset
    - no lemmatization
  - FROLEX morphological dictionary
- experimental (private link)
  - FRO2LEM.PAR http://bit.ly/2QPowIO
    - Simplified Cattex for LGeRM compatibility

https://groupes.renater.fr/wiki/palafra/public/choix_conversions_etiquettes

Textual corpora analysis with Textometry tools & applications

- Content analysis – distant reading applications: linguistics, literary studies, history, philosophy, geography, etc.
- Qualitative tools
  - Kwic concordances of word patterns, text Edition browsing, Progression map
- Quantitative tools
  - Frequency lists, Collocations, N-grams, Keywords, Correspondence analysis…

Computerized methodology



Software to interact with any texts

platform – http://textometrie.org

- 2007-2010 **kick-off** French Research Agency grant
  - members from 6 research laboratories (4 FR, 1 UK, 1 CA)
- 2011-2018 various projects
- **Standard software** architecture (Java+OSGi+J2EE)
  - Key open-source components : R (statistics) & CQP (search engine)
- **Standard corpus formats** :
  - Unicode (TXT), XML, **XML-TEI compatible**
- **Free** & **Open-source** GPL software
  - **TXM Windows**, **Mac OS X** & **Linux** desktop application
  - **TXM Web Portal**: online access (membership, control)
- Efficient : up to 1 billion words

TXM Hands-on

- Download TXM 0.7.9 from http://textometrie.org
- Install TreeTagger for TXM
  - http://txm.sourceforge.net/doc/treetagger/en/install.html
  - Download fro.par and fro2lem.par from
    - http://bfm.ens-lyon.fr/spip.php?article324
    - http://bit.ly/2QPowlO (Huma-Num ShareDocs)
  - Unpack and copy it to TreeTagger/models
- Download and unpack working files
  - http://bit.ly/2VUgSKB (Huma-Num ShareDocs)

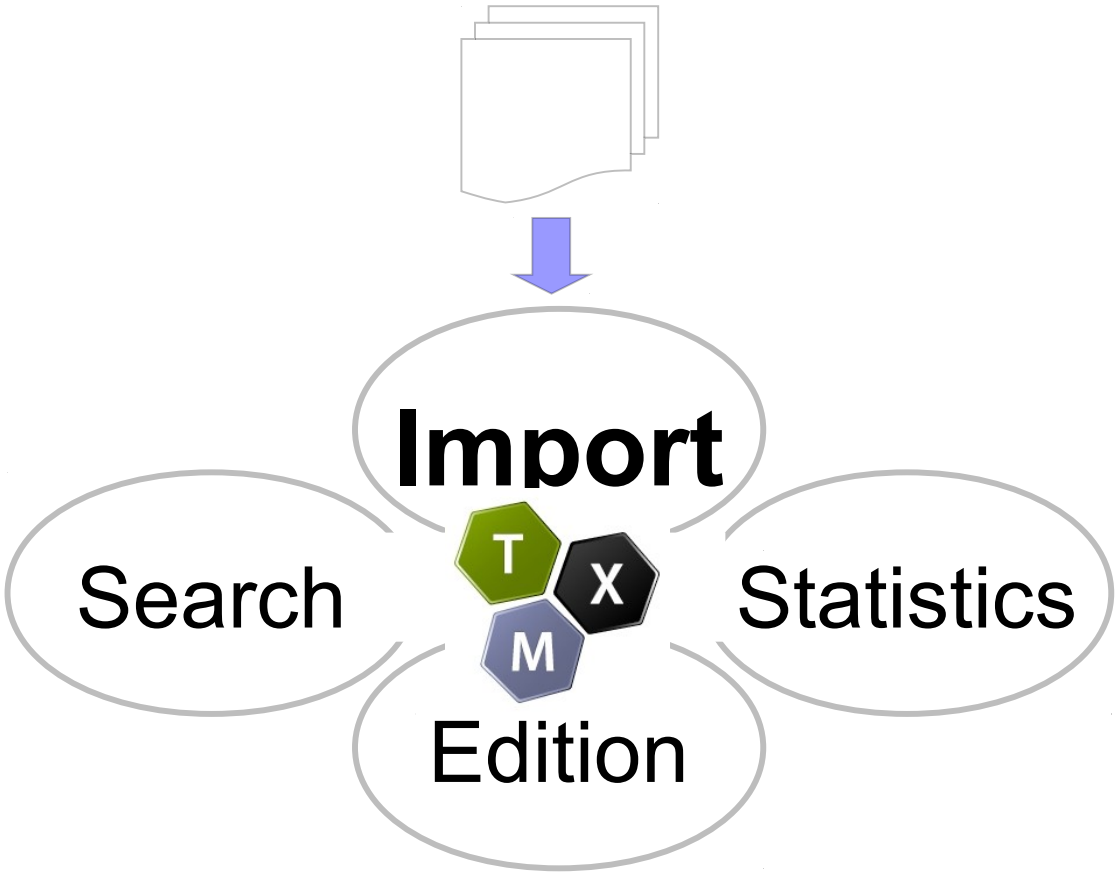
TXM Hands-on

- Working files
  - **Lavrentev_CondorcetWorkshop_2019-01-16.pdf**
    - this presentation
  - CONDORCETEDNOPOS.txm, CONDORCETTRNOPOS.txm and CONDORCETPOS.txm
    - binary corpora (use TXM commands File > Load)
  - Folders **condorcet-ed-nopos**, **condorcet-tr-nopos** and **condorcet-pos**
    - source files for XTZ import module extracted from
      - https://github.com/Jean-Baptiste-Camps/Geste/tree/master/xml

TXM Demo

- Local application or txm.bfm-corpus.org
- *Queste del saint Graal edition*
  - GRAAL corpus
    - browse the synoptic edition
    - lexicon / index / concordance
    - progression (local application only)

TXM import



TXM textual corpus model

- **Textual units** (book, article, interview…)
  Metadata (autor, date, domain, genre…)
  - **Internal structures** (sentence, paragraph, sections…)
    Properties (number, title...)
    - **Lexical units** (words, coumpound words)
      Properties (graphical form, lemma, part of speech…)
  - **Textual planes**
    - Out-of-text (teiHeader, comments…)
    - Speech turn, direct speech...
    - Main language (latin…), Secondary language (Ancient Greek…)
- NLP tools involved (lemmatizers…)
- Text editions for reading and browsing
  - Pagination (page breaks)
  - Rendering (styles), Media (Image, Audio, Video)
- Alignement (parallel corpora)

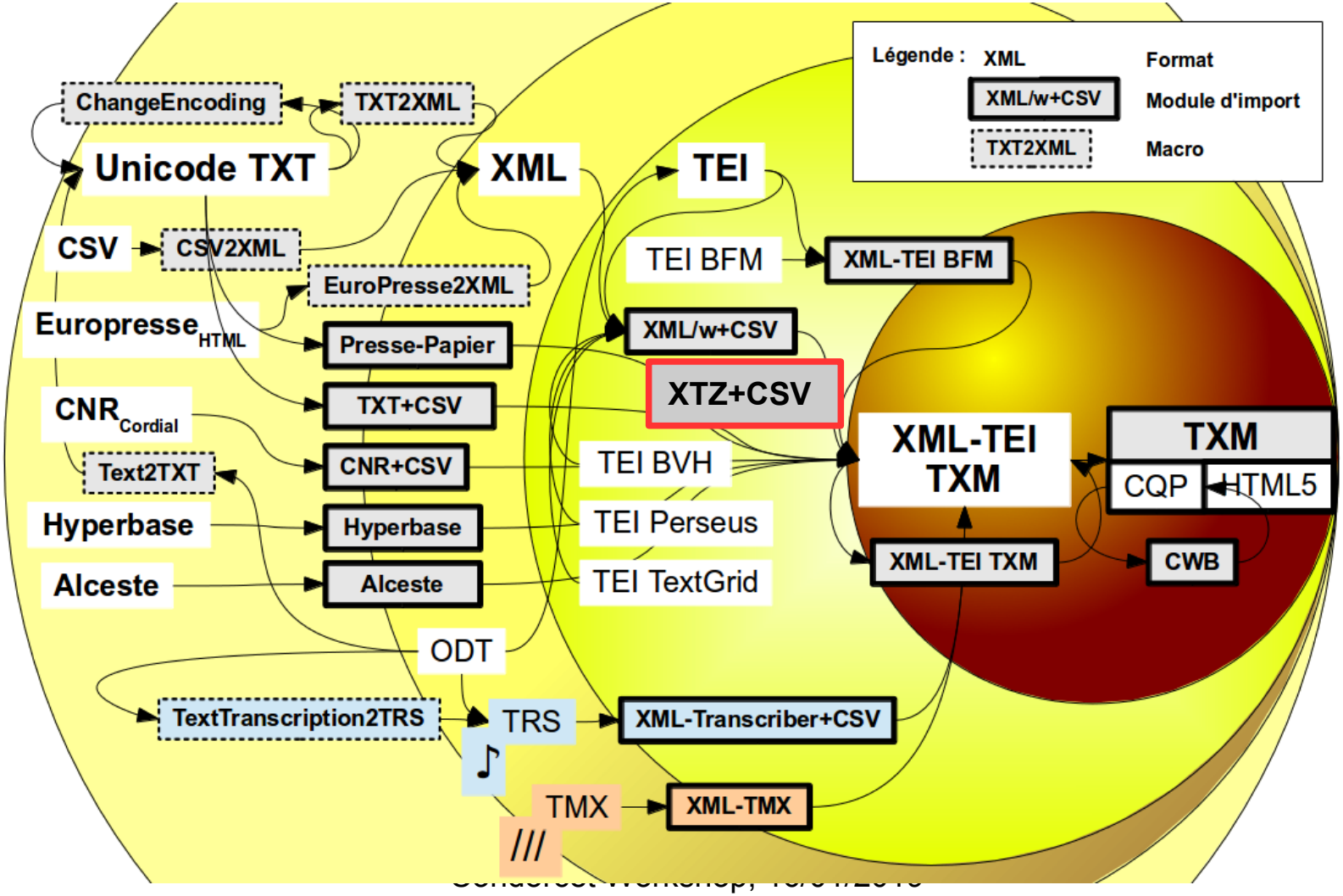
TXM import: 3 types of sources

A) **Written texts** corpora (TXT, XML, XML-TEI) text editions possibly aligned with facsimile images

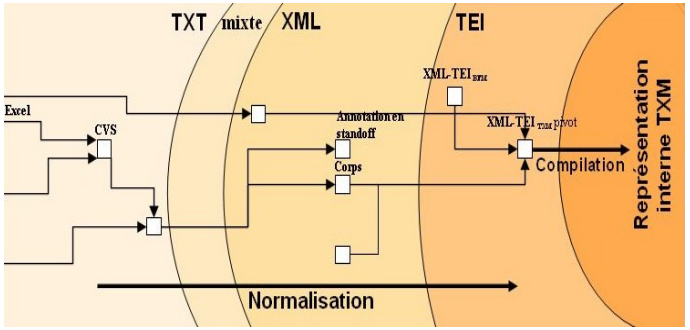
B) **Speech transcription** corpora (XML-TRS), possibly synchronized with audio or video

C) **Multilingual aligned** corpora (XML-TMX), at the a level of a text structure ; paragraph, sentence...

TXM import framework



Import and Analysis Workflow



- directory with TXT files
- XML documents + metadata
- NLP enriched texts
- TEI-TXM XML documents
- Contrasts : sub-corpora & partitions
- Structures
- Lexical facets

TreeTagger

TXM

TXM import modules

corpus input formats

- Various proprietary formats : Hyperbase, Alceste, CNR (Cordial)
- *Calibre* – open ebook digital library (ePub)

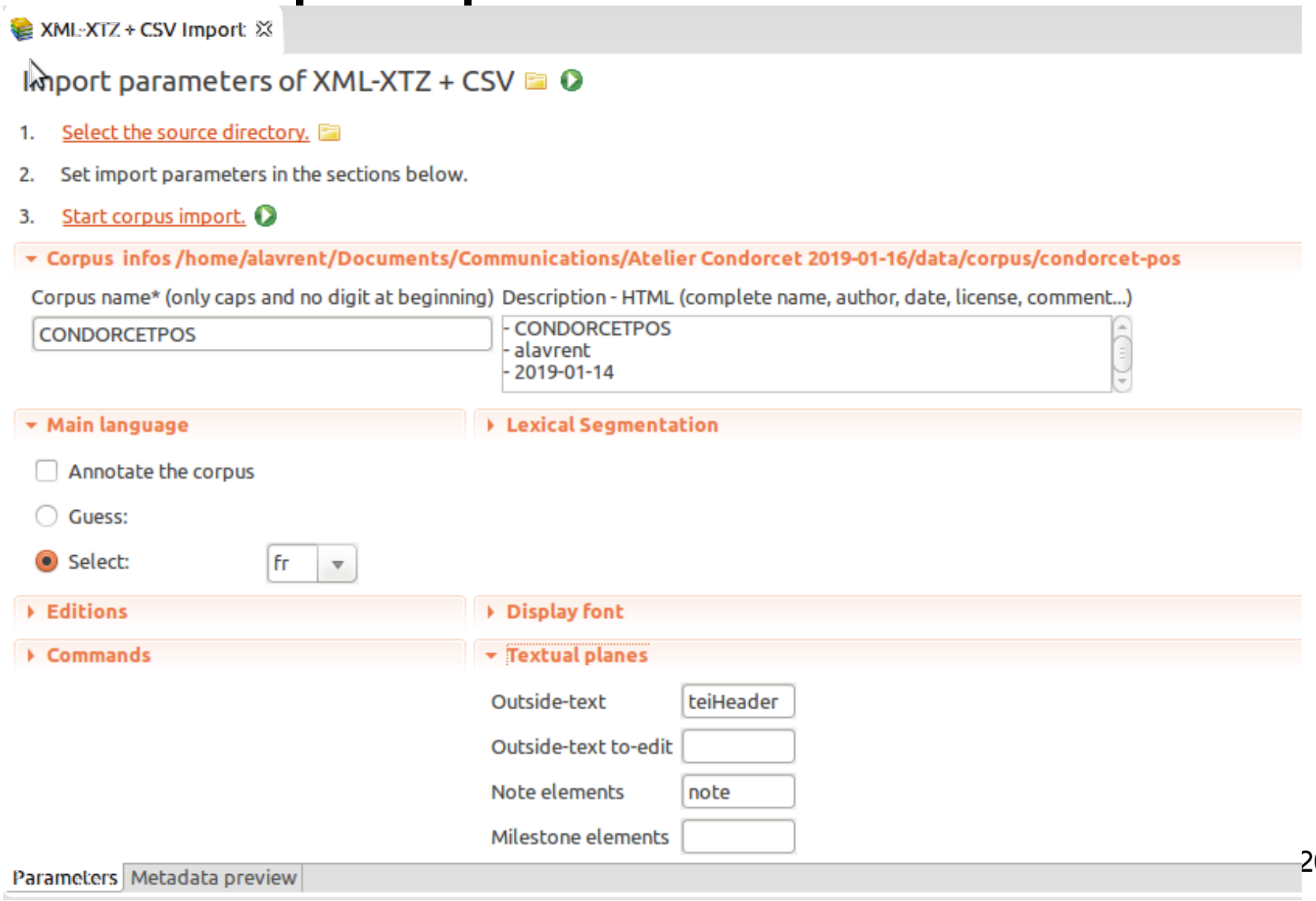
- Copy/Paste
- TXT Unicode+CSV (metadata) : raw texts directory

- XML/w+CSV : XML texts directory
- XML-XTZ+CSV : XML texts directory with XSL scripts

- XML-TEI P5 **BFM** : TEI standard compatible XML
- XML-TEI P5 **BVH / FRANTEXT texts / FRANTEXT search results**
- XML-TEI-TXM : TEI compatible XML+NLP (pivot)

- XML-Transcriber+CSV – audio aligned transcriptions
- *XML-TMX* – multilingual aligned corpora
- *XML-PPS*-Factiva – press portal

TXM Import parameters form



XTZ+CSV Import Module

- TEI Tags Used
  - See TXM User Manual, section 6.1.5 (French)
  - <text> (1 per file)
    - → do not use teiCorpus, group
  - <w> (may be changed in the import form)
  - for the editions: pb, head, p, hi, emph, list, table…
  - the other tags (TEI or not) become structures

XTZ+CSV Import Module

- Customize editions with CSS
  - “css” subfolder
    - [CORPUSNAME].css
    - TXM.css
  - display fonts, page layout, colors, etc.
    - titles, notes, higlights (<hi>, <emph>)
  - all tags of the source document are not available for styling when using “default” edition

XTZ+CSV Import Module

- XSLT transformations
  - 1-split-merge : file reconfiguration
    - currently bugged (TXM 0.7.9)
  - 2-front : preparing for tokenisation
  - 3-posttok :
    - tuning tokenisation,
    - creating references for concordances,
    - projecting word properties
  - 4-edition :
    - customizing the default edition
    - creating additional editions

TXM Data Export

- Tables (index, concordance…)
  - CSV / TSV
- Graphics
  - SVG, JPG, PNG...
- Annotations
  - TEI XML URS, TigerSearch
- Binary corpus
  - TXM = ZIP
    - TEI-TXM XML files

TEI-TXM XML Format

- https://groupes.renater.fr/wiki/txm-info/public/xml_tei_txm
- <tei:w>
  - <txm:form> : the word form
  - <txm:ana> : all annotations, including alternative transcription presentations
    - @type
    - no sub-elements

TEI-TXM XML Format

- <w id="w_qgraal_cm_2643">

  <txm:form>me<ex>n</ex>
  <pb xml:id="page_161v"/>
  <cb xml:id="col_161c"/>
  <lb n="1"/>joient
  </txm:form>

  <txm:ana resp="none" type="#dipl">m‹en›/||ioient</txm:ana>

  <txm:ana resp="none" type="#facs">mē/||ioient</txm:ana>

  <txm:ana resp="none" type="#pos">VERcjg</txm:ana>

  </w>

Thank you!

http://bfm.ens-lyon.fr
http://textometrie.org

textometrie@groupes.renater.fr
bfm@ens-lyon.fr