Improving Access to Text
IMPACT

Technical deliverable documentation

D-EE3.13 Proof of Concept Historical Lexicon for French | EE3

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Gilles Souvay	2011-12-11	Initial version
1.0	Final	"	20 January 2012	Final version delivered

Approvals

Version	Date of approval	Name	Role in project	Signature
0.1	19 January 2012	Katrien Depuydt	WP EE3 leader	OK
1.0	23 March 2012	Max Kaiser	SP EE leader	OK
1.0	23 March 2012	Hildelies Balk	Project Director	OK

Distribution

Version	Date of sending	Name	Role in project
0.1	11 December 2011	Katrien Depuydt	WP EE3 leader
1.0	23 January 2012	Max Kaiser, Hildelies Balk	SP EE leader, Project Director
1.0	6 April 2012	Liina Munari	EC Project Officer

KB
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands

IMPACT – Technical deliverable documentation

D-EE3.13 Proof of Concept Historical Lexicon for French

1. Partner

ATILF

2. Deliverable

D-EE3.13 (French part)

3. Background

The lexica are delivered as an SQL Database and an XML document according to the IMPACT Lexicon Format. The core general lexicon for French is intended to improve both OCR and retrieval for historical French.

The lexica for French rely on the following data:

- Lemma entries in *TLF*: Trésor de La Langue Française, dictionary for modern French[1].
- Lemma entries (when lemma not in TLF) from DMF: Dictionnaire du Moyen Français (1330-1500)[2]
- Modern French morphological lexicon Morphalou[3]
- Words from the IMPACT Ground truth dataset.
- Words from the *Frantext*[4] textual database
- The Middle French knowledge base from the *LGeRM*[5] lemmatiser for middle French

4. Outline of functionality

The *OCR lexicon* has 141,250 entries. The lexicon provides (historical) word form, modern form, lemma, part of speech and frequency. The lexicon was constructed in the following way:

1. From the Morphalou lexicon a "*hypothetical lexicon*" of classical French variant forms was constructed by applying archaization rules.
2. From this hypothetical lexicon, a selection of actually witnessed words was made according the textual corpora mentioned above.

The lexicon contains around 27,800 lemma + part of speech combinations (AVOIR+verbe, AVOIR+subst. are two different entries of TLF). Special lemma names where used for the following categories of words:

NOM_PROPRE for proper name
MOT_ETRANGER for foreign word
MOT_NOMBRE for numbers.

---

[1] www.atilf.fr/tlf

[2] www.atilf.fr/dmf

[3] http://www.cnrtl.fr/lexiques/morphalou/

[4] www.frantext.fr

[5] Souvay G, Pierrel J-M. *LGeRM. Lemmatisation des Mots en Moyen Français. Traitement Automatique des Langues 2009;50:149-72.*

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands

The part of speech set is based on the Morphalou part of speech set:

Abbreviation	*Explanation*
adj.	Adjective
adj. num.	numeral adjective (when not in Morphalou)
adv.	Adverb
interj.	Interjection
loc.	locution, adverb, part of multiword expression, according to Morphalou classification
nom propre	proper name (when not in Morphalou)
subst.	Noun
Verbe	Verb

Frequency is calculated in the IMPACT/ATILF corpus (IMPACT Ground Truth development + subset of Frantext for the 18th century). The IMPACT development GT contains 670,944 tokens (31 379 types).The Frantext subset has 176 texts from 1600 to 1740. They were partially modernised when keyed, 10-20 years ago, at ATILF. The whole corpus contains 10,684,781 tokens (including punctuation), 132,005 types (including words split by line breaks, words, ground truth errors). The first release of this lexicon does not take compound words into account, because it was impossible to distinguish them from words split by line breaks. Another reason is that these entries are omitted in Morphalou lexicon.

The *IR lexicon* has 481,946 entries. 354,341 of them have been provided with context using IMPACT Ground Truth. 127,605 need a further analyse of GT or Frantext corpus. GT errors and printer's errors were corrected in the IR lexicon.

The *historical spelling* of classical French is described by a set of 799 "*archaization"* rules. The rules are based on a rule set defined for the LGeRM lemmatizer developed at ATILF for Middle French, and have been adapted to 17th century French for IMPACT.
Each rule consist of 4 fields:

1. A regular expression pattern matching part of a modern word form
2. The replacement (part of the corresponding historical word form)
3. A condition on the lemma to which the word form belongs, expressed as a regular expression.
4. List of lemma on which the rule may be applied. If empty the rule may be applied to all lemmata..

The rules have been converted and simplified into the IMPACT format by INL for use in the IR experiments..

The *IR evaluation set* consists of 138 pages, 25,273 tokens, 3,781 types from the GT evaluation subset. Each word was lemmatised by LGeRM and checked by human. GT errors and printer's errors were corrected in IR evaluation set.

The following table describes the files included in the deliverable.

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands

File	*Description*
IRLexicon/FrenchIRLexicon.1.0.sql.gz	Impact French IR lexicon, sql dump of database
/IRLexicon/FrenchIRlexicon.1.0.txt	Impact French IR lexicon, text representation. Columns are: 1. word form 2. lemma 3. part of speech 4. modern equivalent word form, 5. quotation 6. <empty column> 7. offset of start of word form in quotation 8. offset of end of word form 9. terminus post quem (year), 10. terminus ante quem 11. title of quoted work 12. author(s) of quote work
IRLexicon/FrenchIRLexicon.tei.1.0.xml.gz	Impact French IR lexicon, TEI XML representation
OCRLexicon/FrenchOCRLexicon.0.9.tf.txt	French OCR lexicon as used in evaluation, based on development set of ground truth
OCRLexicon/OCR_lexicon_v_1_0.tf.txt	French OCR lexicon based on complete ground truth (2 columns: word form and frequency)
OCRLexicon/OCR_lexicon_v_1_0.txt	Version of previous with lemma and part of speech. Columns: historical word form, modern form, lemma, part of speech, frequency
Spelling/RulesForFrench_1_0.xlsx	Rules describing relation between modern and historical (classical) French orthography
IREvaluationSets/Impact_Lemmatization_1_0.zip	IR evaluation set containing 138 pages as TEI XML. The files named <impact id number>.xml contain one page each. The other XML files contain all lemmatized pages for the given work.

5. Evaluation

The technical evaluation of the lexicon has been done by INL. The scientific evaluation is to be found in *Use of computational lexica for OCR and IR on historical documents - a cross-language perspective* (D-EE 2.8).

6. License and IPR protection

All resources produces by ATILF within the IMPACT project are freely available to the research community for non-commercial use. The user has to quote the origin of the resource : ATILF/CNRS & IMPACT project.
Special licences can be agreed on for commercial use.

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands