Technical deliverable documentation

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

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	UA IMPACT Team	07 December 2011	Final version
1.0	Final	“	20 January 2012	Approved 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	12 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

IMPACT – Technical deliverable documentation
D-EE3.13 Proof of Concept Historical Lexicon for Spanish

1. Partner:

University of Alicante (UA)

2. Deliverable

D-EE3.13 Proof-of the concept historical lexica for French, Spanish, Polish, Bulgarian, Slovene and Czech

3. Background

The Spanish Lexicon is intended to improve both OCR and retrieval for historical Spanish documents. The lexicon is delivered as a SQL and XML document according to the IMPACT Lexicon Format. So the dataset provided consists of:

1. Spelling variation rules
2. OCR Lexicon
3. IR Lexicon in SQL and XML format
4. IR Evaluation set

The Spanish Lexicon relies on the following data:

1. The result of corpus-based lexicon building from a selection of texts from Biblioteca Virtual Miguel de Cervantes (BVMC)
2. The result of corpus-based lexicon building from a selection of texts from Biblioteca Nacional de España (BNE)
3. The result of corpus-based lexicon building of the *Diccionario de Autoridades (1726-1739)*
4. Modern Spanish Lexicon data from *Apertium*, used to suggest morphological analyses.

All texts used to build the lexicon are dated between 1499 and 1748.

4. Outline of functionality

The Spanish Lexicon is intended to improve both OCR and retrieval for historical Spanish documents. Of the functionalities described in the description of the Impact Lexicon Structure (D-EE2.1), it supports token-based attestation, dating of attestations, lemma and main part of speech information and also modern equivalents for all word forms described in the lexicon.

The *IR lexicon* contains about 36857) distinct word forms, 11846 lemmata and 613075 attested tokens. The OCR lexicon contains 147192 distinct word forms and 2586242 tokens.

The following simple part of speech set is used for the lemma part of speech information:

N	Noun
Adj	Adjective
Np	Proper Noun
Vblex	Verb
Prn	Pronoun
Num	Numeral
adv	Adverb
det	Determiner
rel	Relative
ij	Interjection
cnj	Conjunction
pr	Preposition
abr	Abbreviation
spc	'Special' used for whitespace
none	Words in other languages

There are no grammatical features (like tense, number …) distinguishing the different words associated with a lemma, but we have added to each word form/token attested a modern equivalent.

When a word is split in parts in the attestation (for instance "al rededor"= alrededor), all parts are assigned to the complete lemma (alrededor), and they are linked by a common group_id in the table wordform_groups.

The opposite case also appears in the lexicon, when a single historical word form represents two modern word forms (for instance "dellos"= de ellos); in that case we have split the word form, using the special tag "EMPTY, spc" when needed, so the analysis looks like: "de, <de>, pr & EMPTY, <EMPTY>, spc & él, <ellos>, prn" (words in angled brackets are the modern equivalents).

The complete lexicon database incorporates material from the following corpora:

corpus_id in lexicon database	*name*	*description*
27	Corpus 11/10/2010	selection of texts from Biblioteca Virtual
45	Traductor	IMPACT ground truth material – development set
39	Development-GT	IMPACT ground truth material – development set
40	Diccionario	*Diccionario de Autoridades (1726-1739)*
41	IR_Evaluation	IMPACT ground truth material – fully lemmatised part of Evaluation subset

The OCR lexicon used in the scientific evaluation is based on all corpora in the database - except for corpus 41 which is also part of the OCR evaluation corpus. The IR lexicon used in evaluation contains material from corpora 27, 45, 39 and 40. A cleaner version of the OCR lexicon could be obtained by removing material from corpus 40. This possibility will be evaluated in the second extension of the IMPACT project.

Short description of the files composing this deliverable:

File	*Description*
./IREvaluationSets/SpanishIREvaluationSet.tei.xml	XML (TEI p5) version of IR evaluation set for Spanish
./IRLexicon/SpanishIRLexicon.v1.sql.gz	SQL dump of Spanish IR lexicon
./IRLexicon/SpanishIRLexicon.v1.tei.xml.gz	XML (TEI p5) version of Spanish IR lexicon
./OCRLexicon/SpanishOCRLexicon_v1.0.txt	OCR lexicon as used in the scientific evaluation
./Spelling/SpellingVariation_spanish.xls	Spelling variation rules

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-EE2.8).

6. License and IPR protection

The licensing follows the consortium agreement.

Product will be integrated as a resource into the OCR workflow, wordlists from the complete historical corpus will be integrated as well into products developed in TR3 and TR5.

The lexicon probably will be made freely available for non-commercial use. However the type of license is still under consideration.