Improving Access to Text
IMPACT

Technical deliverable documentation

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

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Karel Kučera	20 December 2012	Initial version
1.0	Final	"	20 January 2012	Final 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

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 Czech

1. Partner

CUP

2. Deliverable

D-EE3.13 (Czech part)

3. Background

The lexica have been delivered as four SQL databases and XML documents according to the IMPACT format. The lexica reflect four distinct periods of development of modern Czech orthography, with three major spelling reforms implemented in the 19th century (1809, 1843 and 1849).

The four core general lexica are intended to improve both OCR efficiency and information retrieval for 19th-century Czech documents.

The Core General lexica for Czech are based on the following data:

a) dictionaries of J. Dobrovský (published in 1802 and 1821), J. Jungmann (1835 –1839) and F. Š. Kott (1878-1897), with the headwords of these dictionaries expanded into full paradigms;

b) the result of corpus-based lexicon building from 19th-century texts from the text bank of the Czech National Corpus;

c) the result of corpus-based lexicon building from 19th-century texts from selections of the Czech National Library material.

4. Outline of functionality

The four core general lexica of the 19th-century Czech are intended to improve both OCR efficiency and information retrieval for pre-modern Czech documents. Out of the functionalities described in the Impact Lexicon Structure (D-EE.2.1), the four lexica support token-based attestation and dating of attestations, and provide both lemmata and part-of-speech information for all word forms included in the lexica.

In total, the four lexica include 1,327,517 lines, each of them consisting of a word form, its lemma and part of speech information. The lexica cover the periods 1801-1809, 1810-1842, 1843-1849 and 1850-1900, respectively with 321,099, 304,711, 183,079, 518,628 word form-lemma-PoS tuples. Basic information about frequency of word forms, lemmata and parts of speech can be extracted from counting the attestations of these units in texts.

In the four lexica, the following part-of-speech tags are used:

A	Adjective
C	Numeral
D	Adverb
F	foreign word
I	Interjection
J	Conjunction
N	Noun

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

P	Pronoun
R	Preposition
T	Particle
V	Verb
Z	Abbreviation

Apart from the lexica, we deliver:

- A document describing historical spelling variation in Czech
- IR Evaluation sets consisting of manually lemmatized material for about 10.000 tokens selected from the Evaluation subset of the IMPACT ground truth

Short description of the files in this deliverable

- The IR Lexica are delivered as plain text, SQL dump and TEI p5 XML. Attestations (quotations) are provided from the IR evaluation sets.
- The OCR lexica are delivered as plain text files, containing one word form per line
- The IR Evaluation sets are delivered separately as TEI p5 XML only. The SQL dumps would be identical to the dumps of the IR lexica

File	*Description*
IREvaluationSets/LexiconTool_Czech_1800_1809.xml	TEI p5 version of Czech IR evaluation set, 1800-1809
IREvaluationSets/LexiconTool_Czech_1810_1842.xml	TEI p5 version of Czech IR evaluation set, 1810-1842
IREvaluationSets/LexiconTool_Czech_1843_1849.xml	TEI p5 version of Czech IR evaluation set, 1843-1849
IREvaluationSets/LexiconTool_Czech_1850.xml	TEI p5 version of Czech IR evaluation set, 1850-onward
IRLexica/CzechIRLexicon_1800_1809.utf8.txt	Plain text version of Czech IR Lexicon, 1800-1809
IRLexica/CzechIRLexicon_1800_1809.xml	TEI p5 version of Czech IR Lexicon, 1800-1809
IRLexica/CzechIRLexicon_1810_1842.utf8.txt	Plain text version of Czech IR Lexicon, 1810-1842
IRLexica/CzechIRLexicon_1810_1842.xml	TEI p5 version of Czech IR Lexicon, 1810-1842
IRLexica/CzechIRLexicon_1843_1849.utf8.txt	Plain text version of Czech IR Lexicon, 1843-1849
IRLexica/CzechIRLexicon_1843_1849.xml	TEI p5 version of Czech IR Lexicon, 1843-1849
IRLexica/CzechIRLexicon_1850_.utf8.txt	Plain text version of Czech IR Lexicon, 1850-onward
IRLexica/CzechIRLexicon_1850_.xml	TEI p5 version of Czech IR Lexicon, 1850-onward
IRLexica/LexiconTool_Czech_1800_1809.sql	SQL dump of Czech IR Lexicon, 1800-1809
IRLexica/LexiconTool_Czech_1810_1842.sql	SQL dump of Czech IR Lexicon, 1810-1842
IRLexica/LexiconTool_Czech_1843_1849.sql	SQL dump of Czech IR Lexicon, 1843-1849
IRLexica/LexiconTool_Czech_1850.sql	SQL dump of Czech IR Lexicon, 1850-onward
OCRLexica/CzechOCRLexicon_1800_1809.txt	Czech OCR Lexicon, 1800-1809
OCRLexica/CzechOCRLexicon_1810_1842.txt	Czech OCR Lexicon, 1810-1842
OCRLexica/CzechOCRLexicon_1843_1849.txt	Czech OCR Lexicon, 1843-1849
OCRLexica/CzechOCRLexicon_1850_1900.txt	Czech OCR Lexicon, 1850-onward
Spelling/SpellingVariations.docx	Document describing historical spelling variation in 19th century Czech

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

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). Preliminary experiments have shown good coverage. For further evaluation experiments, ground truth data have been developed as part of the project.

6. License and IPR protection

All resources produces by CUP 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 : CUP & IMPACT project. Special licences can be agreed on for commercial use.