Technical deliverable documentation

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

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Final	Stoyan Mihov, BAS	12 December 2011	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	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 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-EE.3.13 Proof of Concept Historical Lexicon for Slovene

1. Partner

IIKT – BAS

2. Deliverable D-EE.3.13 Bulgarian Lexica

This deliverable consist of 3 parts:

a. Bulgarian core lexicon
b. Pattern set for Bulgarian
c. IR Evaluation set for Bulgarian

3. Background

a. The **Bulgarian core lexicon** is delivered as an SQL Database and an XML document according to the IMPACT Lexicon Format. The core general lexicon for Bulgarian is intended both for the improvement of OCR and information retrieval for late 19th century Bulgarian. This is the first release of the lexicon. Lexicon development will continue until M48. Lexicon content of Bulgarian will be delivered to WP-TR-5 (Language modelling and dictionaries in OCR) and WP-TR-3 (Adaptive OCR) to be used for the development of their tools.
The Core General lexicon for Bulgarian is a result from a corpus-based lexicon building from a selection of the Bulgarian National Library OCR-ed material.
In the first release we deliver data for all the words with more than one occurrence in the corpus. Thus the most frequent variations of words are captured.

b. The **Pattern set for Bulgarian** is delivered in UTF-8 text format (see the Appendix). Each line describes a historical variation pattern in the format:

```
[@]<modern>[@]:[@]<historical>[@]
```

Here the metasymbol @ indicates that the pattern refers only to prefix/suffix depending on the position of the symbol with respect to the sequence of characters. This set describes the regular variations in the Bulgarian language established during the work on an OCR-corpus of the late 19th materials of the Bulgarian National Library, which was also used for the means of the Lexicon building.

The pattern set for Bulgarian will be delivered to WP-TR-5 (Language modelling and dictionaries in OCR) and WP-TR-3 (Adaptive OCR) to be used for the development of their tools.

c. The **IR Evaluation set for Bulgarian** is delivered in twelve UTF-8 text files. Each of the files is named: <file_name>.tag and corresponds to the file: <file_name>.xml from the GT Evaluation Set for Bulgarian uploaded on the Prima server at USAL. The tokens in the XML-file are described in order on separate lines in the tagged file as follows:

```
<historical_word>\t<modern_lemma>:<modern_word_form>{|<modern_lemma>:<modern_word_form>}
```

The files were selected as to cover all the sources provided by the National Library of Bulgaria for the aims of the project. The IR Evaluation set for Bulgarian is delivered to delivered to WP-TR-5 (Language modelling and dictionaries in OCR) and WP-TR-3 (Adaptive OCR) to be used for the development of their tools.

4. Outline of functionality

a. The **Bulgarian core lexicon** is intended to improve both OCR and retrieval for historical Bulgarian documents where the modern font is used. Of the functionalities described in the description of the Impact lexicon structure (D-EE.2.1), it supports token-based attestation, dating of attestations, and lemma and part of speech for all word forms described in the lexicon. The structure of the IR lexicon for Bulgarian allows to easily extract the historical words used in the late 19th century by taking the list of entries in the lexicon and ignoring the rest information. In this way an OCR lexicon for Bulgarian is implicitly contained in the IR lexicon.

Currently there are 29 796 distinct word forms, 12 436 distinct lemmata and 32 947 distinct lemma/wordform combinations in the lexicon. Frequency information can be extracted by counting the attestations for word forms and/or lemma. Although it does not provide a precise measure for the usage of this particular word form it is sufficient to recognize the frequently from infrequently used words.

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

Г	**Verb**
М	**Pronoun**
МЕЖ	**Interjection**
НАР	**Adverb**
ПРЕД	**Preposition**
ПРИ	**Adjective**
С	**Noun**
СЮ	**Conjunction**
ЧА	**Particle**
Ч	**Numeral**

The problems arising from the OCR-errors in the corpus were handled by automatically detecting some of them and manually verifying them. In this way the erroneously recognized words were excluded from consideration. Furthermore during the attestation process the suspicious words were manually compared with the image in order to guarantee that the chosen example correctly illustrates the meaning of the word.

b. The **Pattern set for Bulgarian** is intended to improve the OCR and retrieval for late 19th century Bulgarian language. Based on the pattern set one can automatically generate possible descendants of a historical word into the modern language. Guided by a modern dictionary one can further select the plausible ones. The reverse operation is feasible, i.e. starting with a modern word one can imagine all different ways of its transcription in the late 19th century which is needed for the retrieval.

The pattern set was completed in two stages. In the first stage 8 expert patterns were determined based on the knowledge about the late 19th century Bulgarian on general. During the work on the corpus and the building of the lexicon other regular transformations were detected and suggested by the experts. These were automatically applied on the entire corpus and the experts again verified the resulting pairs. In this way the patterns, which described best the evolution of the language were preserved and the rest were neglected.

c. The **IR Evaluation set for Bulgarian** is intended to support the retrieval for late 19th century Bulgarian language. It provides complete information for modern lemma(ta) and modern word form(s) for 13 437 historical words in the frame of running text. It also reflects the ambiguities where one historical word may refer to one or more modern lemmata or modern word forms.
In order to facilitate the annotation process the selected documents were first automatically preprocessed and candidates for annotation were attached to each token from the files. Afterwards the annotation was verified and corrected manually.

Short description of the files in this deliverable:

File	*Description*
./IRLexicon/BulgarianIRLexicon.v1.0.xml	“Lextractor” XML for the Bulgarian IR lexicon
./IRLexicon/BulgarianIRLexicon.v1.0.sql	(My)sql dump of the Bulgarian IR lexicon in impact database format
./IRLexicon/BulgarianIRLexicon.v1.0.tei.xml	TEI p5 version of the Bulgarian IR lexicon
./OCRLexicon/developmentSet.tf	Type-frequency list of the development part of the IMPACT ground truth as used in the first OCR experiments for Bulgarian
IREvaluationSet/BulgarianIREvaluationSet.zip	IR Evaluation Set

5. Evaluation

The technical evaluation of the Bulgarian lexica 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

The licensing follows the consortium agreement.

The product will be integrated as a linguistic resource into the OCR workflow, word lists from the complete historical corpus will be integrated as well into products developed in TR3 and TR5.

So far no resources available by IMPACT are used for the development of the product.

Appendix. List of Patterns for Bulgarian

е:ѣ
я:ѣ
ъ:ѫ
я:ѭ
@:ъ@
@:ь@
и:ы
и:і
н:нн
с:сс
м:мм
л:лл
е:ъе
ия@:ий@
та@:ьта@
а@:ѫ@
ат@:ѫтъ@
ят@:ѫтъ@
е:ье
жн:ждн
не@:ние@
нето@:нието@
@из:@ис
ият@:ийтъ@
раз:рас
ъ:ь
ме@:ми@
им@:име@
ем@:еме@
ър:ръ
ръ:ър
ъл:лъ
лъ:ъл
к:гк
ят@:атъ@
ал:ѫл
р:рр
аещ:ающ
уващ:ующ