Improving Access to Text
IMPACT

Technical deliverable documentation

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

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Janusz Bien	11 December 2011	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 Polish

1. Partner

UWAR

2. Deliverable

D-EE3.13 (Polish part)

3. Background

The electronic resources concerning Polish historical texts are, with a one notable exception, practically nonexistent, as the researches still concentrate on printed transcribed critical editions. The exception is so called Late Middle Polish Dictionary, the official name being "the dictionary of Polish language of 17th and the first half of 18th century". This is a work in progress being prepared in the Institute of Polish Language of Polish Academy of Sciences, since 2004 developed in the electronic form, maintained as a Postgress database and published in the form of continuously updated web site (http://sxvii.pl/). A snapshot of the database has been kindly provided by the Institute for the use in the IMPACT project. The snapshot contained in particular 13 556 entries in various stages of elaboration (only 214 considered finished), 30 193 inflexional forms and 44 861 quotations from 761 sources. Inflexional forms are in modern spelling, assigned to the appropriate lemmata and tagged with the grammatical categories. Some of the sources of quotation are late editions with modernized spelling and some are the original works; in consequence some quotations are in modern spelling and some in the spelling close to the original one (for technical reasons ligatures has been replaced by appropriate letters etc.). The quotations has been processed using INL Attestation Tool.

As Polish language, like other Slavonic language, has rich inflexion, there has been several approaches to the problem of automatic morphological analysis of contemporary Polish texts. For the IMPACT project two of such resources has been used. The first one is the SAM morphological analyser by Krzysztof Szafran which is based on the so called "Schematic index a tergo of Polish wordforms" (cf. e.g. http://bc.klf.uw.edu.pl/87/). Optionally the analyser can consult the entry list of so called Early Modern Polish dictionary (official title just "Dictionary of Polish language", edited by W. Doroszewski) which covers the period since the second half of 18th century to the first half of 20th century. The second resource used is the relatively recent Morfeusz SGJP morphological analyser, which provides in particular 5 086 141 inflexional forms with grammatical information.

A very important source of information about historical Polish language is the monumental dictionary of Samuel Linde, published in 1807-1814, as it was the first Polish dictionary aiming at documenting the language actually used in texts. Using the dictionary directly was not possible for technical and economical reasons (the dictionary is extremely difficult to OCR), so we focused on an index a tergo published in 1965. Digitalizing it seemed an easy task but it was not. Nevertheless after semi-automatic validation and manual proof-reading a quite reliable list of about 80 000 of entries has been obtained, which was supplemented by grammatical information using the analysers mentioned above.

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

We have been using also a resource specific to the IMPACT project, namely the ground-truth texts, which unfortunately became available quite late. The first part of GT consists of 4094 pages from 8 volumes (including a famous and important 18th century encyclopedia in 4 parts) published in years 1617-1756. The second part consists of 599 pages coming from 25 news pamphlets with the size ranging from 6 to 32 pages published in years 1570 to 1728. For IR evaluation purposes a subset of the text has been processed with INL Lexicon Tool.

It should be noted that many texts are printed in Fracture fonts (extended of course with additional letters needed for Polish) while using Roman font for Latin quotations.

Using for historical data the morphological analysers designed for modern texts required taking into account the spelling variations and getting rid of ligatures. A simple formalism based on regular expressions has been designed and a set of over hundred rules has been prepared. Many of them are quite obvious, e.g. replacing the old form of letters like long s by modern ones. Some of them required good insight into the history of Polish language, they have been written by the historical linguist being a member of the team. The rules took into account the scarce literature on the subject and has been tested on the GT data.

4. Outline of functionality

The *OCR lexicon* consists of about 112.000 words from the quotations in the Late Middle Polish dictionary and over 200 000 words from all the GT texts.

File	*Description*
OCRLexicon/IMPACT_PSNC_GT_final_freqlist_ligs.txt	Frequency list of IMPACT ground truth material for Polish
OCRLexicon/PolishOcrLexicon.evaluationVersion.txt	Polish OCR lexicon, version used in the evaluation
quotations_17v.tf	Frequency list of words in quotation text in the dictionary of Polish language of 17th and the first half of 18th century

The *main IR lexicon* based on the dictionary of the XVII century, converted to impact database structure and also as an XML export from that database.

File	*Description*
IRLexicon/polishAttestationLexicon.tei.xml.gz	XML (TEI p5) version of the polish IR lexicon
IRLexicon/polishAttestationLexicon.sql.gz	SQL dump of the IR lexicon database

The *IR evaluation sets* are databases in impact Lexicon consisting of a full lemmatization of about 10.000 tokens chosen by a random selection from the evaluation part of the IMPACT polish ground truth set.

File	*Description*
PolishIRSet1.sql	Random choice of pages from GT set (emphasizing material before 1700)
PolishIRSet2.sql	Random choice of pages from GT set (material from *Nowe Ateny*, 18th century)
PolishIRSet1.tei.xml	XML version of PolishIRSet1.sql
PolishIRSet2.tei.xml	XML version of PolishIRSet2.sql

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 *spelling variation rules* (subdirectory Spelling) consist of the rules proper in the file podstawienia2.csv and the exceptions in the file stoplisty1.txt. The syntax is described briefly in the file rules_syntax.txt. For testing purposes a Python script is provided in the file normalize.py. It should be used in the following way:

Usage: normalize.py [OPTIONS] RULES_FILE INPUT_FILE OUTPUT_FILE

Options:

```
 --version             show program's version number and exit
 -h, --help            show this help message and exit
 -f, --frequency-list  input file is frequency list
 -v, --verbose         print information about normalized words
 -e EXCEPTIONS, --exceptions=EXCEPTIONS
                       exception file
 -s SEPARATOR, --separator=SEPARATOR
                       CSV separator
```

Material pertaining to the index of the Linde dictionary

The main file containing the digitized form of the index a tergo of Linde dictionary entries is Linde/*iLinde_f2.txt*. The file format is documented in *Documentation/ iLinde_format.txt.*

The Linde's dictionary entries, with the part of speech information and other grammatical information added automatically. The file Linde/iLinde_Morfeusz.txt contains 40 997 entries recognized by Morfeusz analyzer; although oriented towards the modern vocabulary, this analyser has a very good coverage of proper names. The file Linde/Linde_SAManal.txt contains 36 554 entries recognized by the SAM analyser working in so called exact mode, i.e. accepting only the lemma attested by the Late Modern Polish dictionary (the dictionary does not contain proper names). The grammatical information is in a compact format used in the printed version of "Schematic index a tergo of Polish wordforms". The file Linde/iLinde_SAMguesses.txt contains hypothetical grammatical information based on the rules of the schematic index provide for almost 40 000 entries.

Information about the part of speech encoding in the SAM output can be found in the report Documentation/ tr226.pdf.

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

The intention is to make all data freely available, but for distributing the resources derived from the Late Middle Polish dictionary the explicit permission of the Institute of Polish Language of Polish Academy of Sciences should be obtained, as the current agreement covers only their use in the IMPACT project. Distribution of resources derived from Morfeusz and SAM analyser should adhere respectively to their licenses.

The spelling variation rules resources should be considered as copyrighted by the University of Warsaw and released on the GNU General Public License.

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