Improving Access to Text
IMPACT

Technical deliverable documentation

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

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Tomaž Erjavec, JSI	11 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-EE3.13 Proof of Concept Historical Lexicon for Slovene

1. Partner

JSI

2. Deliverable

D-EE3.13 (Slovene part)

3. Background

The proof of concept historical lexicon for Slovene is as a tabular file, appropriate for simple conversion into other formats, such as a SQL Database and as XML encoded lexicon, structured according to the Text Encoding Initiative Guidelines P5. It is intended to improve both OCR and retrieval for historical Slovene documents.

The Slovene lexicon has been automatically extracted from the “goo300k” hand-annotated corpus, compiled in the scope of IMPACT with additional funding from the Google Award “Developing computational models of historical Slovene”, received jointly by JSI and the Scientific research centre of the Slovenian Academy of Sciences and Arts.

The corpus text basis was compiled in June 2011, where only a portion of the Ground Truth Dataset from the National University Library (NUK) of Slovenia was available, so the corpus was supplemented by additional materials. The following sources of proof-read historical texts, together with their facsimiles, were the basis for the corpus:

1. Successive selected pages from three religious books, from the end of the 16th, 17th and 18th centuries respectively. The scans of the books and raw OCR were provided by the Scientific Research Centre of the Slovenian Academy of Sciences and Arts. The pages of the first two of these books also represent the oldest material in the corpus, barely comprehensible to today's speakers.
2. Complete books from the second half of the 18th and first half of the 19th century. The scans and proof-read transcriptions were provided by NUK in the scope of IMPACT. The books were encoded in PageXM, a format specially designed to facilitate the development of OCR software. The books were written in Slovenian, and span religious books, plays, fiction and even a cookbook. Difficult to understand by today's speakers.
3. Selected issues of one Slovenian newspaper, first published in 1843, and continuing to 1890. The facsimiles and transcriptions were also provided by NUK in PageXML.
4. The AHLib digital library, containing complete books, mostly from the second half of the 19th century. The books are translations of German books, and span a wide variety of topics, from fiction, to text-books on various subjects. The library was proof-read and marked up in TEI in the scope of a project by the Austrian Academy of Sciences, in which JSI also collaborated. This part of the text collection was by far the largest, containing about 70 books. The text is in general easy to understand, but contains many spelling changes to today's norm, degrading the performance of HLT tools trained on corpora / lexica of contemporary Slovene.

To arrive at the goo300k corpus, a sampling procedure was developed, which takes the individual pages from the collection and selects random pages, but ensuring a rich mix of text types and balanced coverage over time. Here, however, more weight was given to younger materials, as the main focus of the corpus is in providing HLT support for historical language, and the language of the 19th century is still similar enough to the contemporary one for such methods to yield good results, as well as being the most useful, as there is orders of magnitude more text available from the 19th century than from earlier times.

Table 1 gives the size of the according to the time periods, and overall, by the number of units (book or newspaper samples), the number of pages (the individual unit of sampling), and the approximate number of tokens. The set size of the corpus was 1,000 pages, which was estimated to be the right size for the manual annotation to be feasible given the financial and time constraints of the project.

Period	Units	Pages	Tokens
1584	1	8	6000
1695	1	27	10000
1751-1800	8	155	27000
1801-1850	12	206	74000
1851-1875	36	380	126000
1876-1900	23	224	51000
∑	81	1000	296000

Table 1. goo300k corpus size by time period

The corpus was first automatically annotated, using the ToTrTaLe tool (developed in the scope of Impact, in 2010) which tokenises the text, sentence segments it, transcribes historical words to their contemporary form, tags it with morphosyntactic descriptions and lemmatises it. For tagging and lemmatisation the tool uses models trained on contemporary Slovene, so the transcription step is not only useful by itself, but also crucial for relatively good tagging and lemmatisation. The transcriptions is operationalised by the Vaam (Variant Approximate Matching) finite-state library, developed by LMU, which uses a lexicon of modern word-forms and a set of transcription patters of typical spelling changes that associate historical words to contemporary ones.

By inspecting the unannotated corpus we first developed a set of transcription patterns, and them, with the help of the LeXtractor editor, also developed by LMU assigned contemporary word-forms to the most frequent (and, typically, unpredictable) words in the collection. With this static lexicon and transcription patterns we then automatically annotated the corpus. The corpus, as well as the complete text collection, which was also automatically annotated, was also mounted under a Web-based concordancer with CQP as its back-end.

In the second step the automatically assigned annotations were manually checked and corrected. A team of annotators, most of them students involved in previous annotation project were hired, while the oldest three books were annotated by PhD students in historical Slovenian. The annotation editor used was Cobalt, developed by INL, and its user manual was adapted for Slovene, and additional reference materials (Annotator's Cookbook, FAQ) were written in tandem with training the annotators on sandbox corpora.

The hand-corrected corpus was regularly mounted on the Web concordancer, which provides searching and displaying over all layers of token annotation, including the name of the annotator and time of validation.

The goo300k corpus was annotated with a view to extracting from it the IMPACT lexicon for Slovene, which would be an interesting resource for humans, but also for HLT development, in particular, as the resource for building a good model of historical Slovene for ToTrTaLe. Therefore attention was given to both aspects; on the digital dictionary side, extinct words were given glosses with their closest contemporary equivalents and the source of this ``translation''; on the computational lexicon side, historical/modern word boundaries are carefully brought into correspondence (tokenisation), abbreviations (sentence segmentation) and foreign passages (tagging and lemmatisation) are identified, as are typos in the source.

The manual annotation, proceeding from June to December then corrected mistakes in the transcriptions and tokenisation, contemporary word-form equivalents, PoS tags, and lemmas, and possibly added glosses to extinct words.

The lexicon export from the corpus was performed in two ways. In consultation with JSI as regards the format, INL developed a procedure, which dumps the corpus from Cobalt into XML format compliant with the Text Encoding Initiative Guidelines, TEI P5. This corpus was first slightly post-processed, and them a procedure was developed, which extracts from the lexicon in a tabular format, giving all the annotations for each entry, together with all the bibliographical references from which the attestations are taken. This format is simple for processing, but does not contain the concordances (attestations) from the corpus. More recently, again in consultation with JSI as to the format, INL developed a procedure to dump the Cobalt-internal lexicon directly into XML, again compliant with TEI P5. This procedure is more involved, as it hierarchically groups the entries according to lemma, then modern word-form, then normalised word-form, and finally the exact citation word-form, and gives the attestations for each citation word-form. This lexicon is still appropriate for importing into tools, but is more human readable – we have also developed a conversion from the TEI XML lexicon into a HTML format, appropriate for browsing.

4. Outline of functionality

The Slovene lexicon is intended both to improve OCR and retrieval of historical Slovene documents. It supports token-based attestations, dating of attestations, modern-form, lemma and lexical part-of-speech information.

The (tabular) lexicon has almost 60,000 entries, 55,000 different citation word-forms, 47,000 normalised word-forms (lower cased and with vowel diacritics removed), 37,000 modernised word-forms and 19,000 lemmas. Just under half of the entries belong to historical words, i.e. words where the normalised form is different form the modernised form.

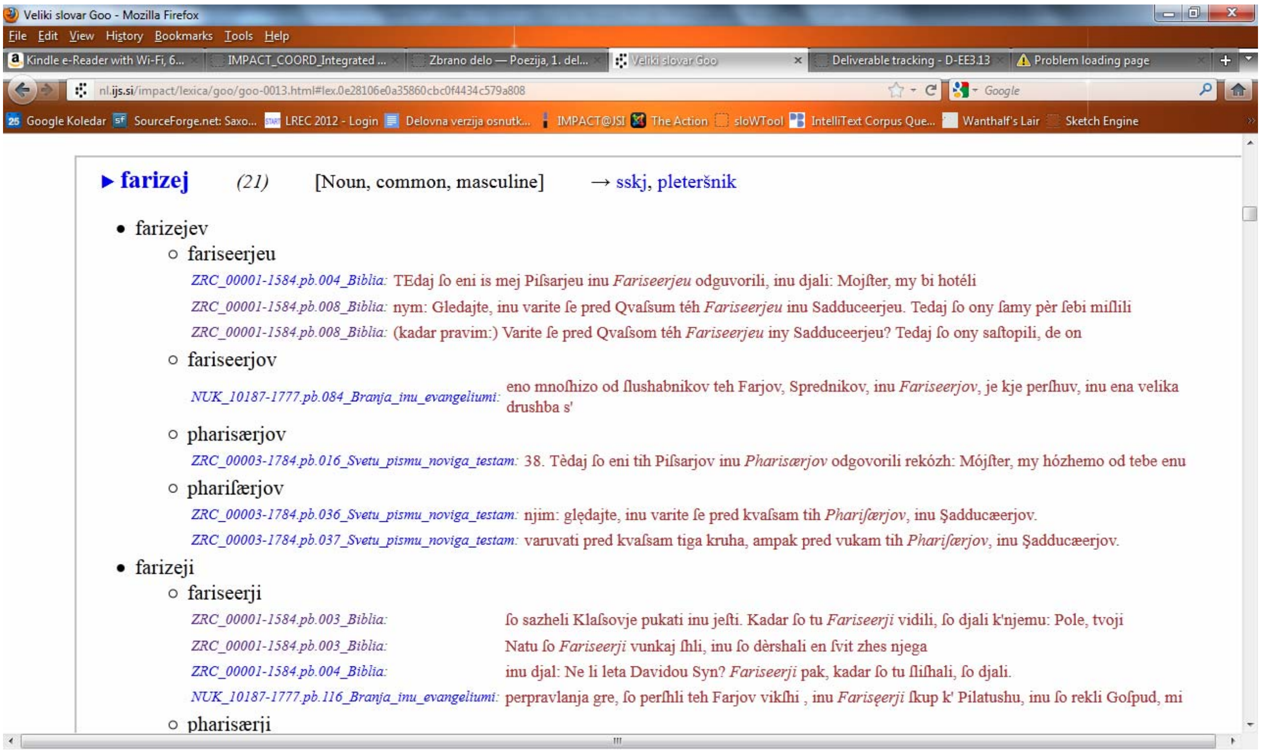
The part-of-speech information included in the corpus and lexicon follows the JOS morphosyntactic specifications for Slovene (http://nl.ijs.si/jos/josMSD-en.html) but all the inflectional information has been excluded, retaining only a coarse-grained tagset, covering mostly just lexical (lemma) features of words, and comprising 32 tags. The specifications for this tagset are given at http://nl.ijs.si/imp/msd/html-en/.

To give an idea of the kind of information included in the (TEI) lexicon, we give in Figure 1 a screenshot of some entries presented in the HTML visualisation of the lexicon.

Output	Delivery method
Documentation	PDF file
Lexicon Data	XML TEI export and SQL dump of lexicon database

Brief description of the files in this deliverable:

File	*Description*
IREvaluationSets/SloveneIREvaluationSet.zip	The IR evaluation set, consisting of 152 pages from the IMPACT evaluation ground truth
IRLexicon/LexiconTool_Goo.2011-12-15.sql.gz	SQL dump of Slovene lexicon database
IRLexicon/Lexicon_Goo.2011-12-14.tei.xml.gz	TEI P5 export of Slovene lexicon database
OCRLexicon/SloveneOCRLexicon.1.0.txt	Type frequency list used as OCR lexicon in the scrientific evaluation
Spelling/example_data_spelling_variation.txt	Example dataset for historical spelling variation taken from the non-evaluation portion of the corpus
Spelling/gooLex.extractedPatterns.txt	Raw extraction of historical variation patterns from the example dataset


Veliki slovar Goo - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Kindle e-Reader with Wi-Fi, 6...
IMPACT_COORD_Integrated ...
Zbrano delo — Poezija, 1. del...
Veliki slovar Goo
Deliverable tracking - D-EE3.13
Problem loading page
nl.ijs.si/impact/lexica/goo/goo-0013.html#lex.0e28106e0a35860cbc0f4434c579a808
Google
Google Koledar
SourceForge.net: Saxo...
LREC 2012 - Login
Delovna verzija osnutk...
IMPACT@JSI
The Action
sloWTool
IntelliText Corpus Que...
Wanthalf's Lair
Sketch Engine

▸ **farizej** *(21)* [Noun, common, masculine] → sskj, pleteršnik

- farizejev
  - fariseerjeu
    - *ZRC_00001-1584.pb.004_Biblia:* TEdaj ſo eni is mej Piſsarjeu inu *Fariseerjeu* odguvorili, inu djali: Mojſter, my bi hotéli
    - *ZRC_00001-1584.pb.008_Biblia:* nym: Gledajte, inu varite ſe pred Qvaſsum téh *Fariseerjeu* inu Sadduceerjeu. Tedaj ſo ony ſamy pèr ſebi miſlili
    - *ZRC_00001-1584.pb.008_Biblia:* (kadar pravim:) Varite ſe pred Qvaſsom téh *Fariseerjeu* iny Sadduceerjeu? Tedaj ſo ony saſtopili, de on
  - fariseerjov
    - *NUK_10187-1777.pb.084_Branja_inu_evangeliumi:* eno mnoſhizo od ſlushabnikov teh Farjov, Sprednikov, inu *Fariseerjov*, je kje perſhuv, inu ena velika drushba s'
  - pharisærjov
    - *ZRC_00003-1784.pb.016_Svetu_pismu_noviga_testam:* 38. Tèdaj ſo eni tih Piſsarjov inu *Pharisærjov* odgovorili rekózh: Mójſter, my hózhemo od tebe enu
  - phariſærjov
    - *ZRC_00003-1784.pb.036_Svetu_pismu_noviga_testam:* njim: ględajte, inu varite ſe pred kvaſsam tih *Phariſærjov*, inu Şadducæerjov.
    - *ZRC_00003-1784.pb.037_Svetu_pismu_noviga_testam:* varuvati pred kvaſsam tiga kruha, ampak pred vukam tih *Phariſærjov*, inu Şadducæerjov.
- farizeji
  - fariseerji
    - *ZRC_00001-1584.pb.003_Biblia:* ſo sazheli Klaſsovje pukati inu jeſti. Kadar ſo tu *Fariseerji* vidili, ſo djali k'njemu: Pole, tvoji
    - *ZRC_00001-1584.pb.003_Biblia:* Natu ſo *Fariseerji* vunkaj ſhli, inu ſo dèrshali en ſvit zhes njega
    - *ZRC_00001-1584.pb.004_Biblia:* inu djal: Ne li leta Davidou Syn? *Fariseerji* pak, kadar ſo tu ſliſhali, ſo djali.
    - *NUK_10187-1777.pb.116_Branja_inu_evangeliumi:* perpravlanja gre, ſo perſhli teh Farjov vikſhi , inu *Farisęerji* ſkup k' Pilatushu, inu ſo rekli Goſpud, mi
  - pharisærji

Figure 1. Example entry from the HTML rendering of the TEI P5 Slovene historical lexicon.

5. Evaluation

The corpus, on which the lexicon is based has been thoroughly checked for errors, in a variety of ways – through Cobalt and comparison with the facsimiles, through the dedicated concordancer, and by exporting the modernised version of the texts and reading though those to notice any inconsistencies. The lexicon has been checked formally, by using an XML validator, with other ad-hoc validation scripts, and by reading the lexicon in its HTML format.

Preliminary experiments by INL have shown good coverage and significantly better IR with using the Slovene historical lexicon.

6. License and IPR protection

Licencing follows the consortium agreement. The lexicon will also be made available via the Creative Commons – Attribution licence.