The goo300k corpus of historical Slovene

Tomaž Erjavec

Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

The paper presents a gold-standard reference corpus of historical Slovene containing 1,000 sampled pages from over 80 texts, which were, for the most part, written between 1750 – 1900. Each page of the transcription has an associated facsimile and the words in the texts have been manually annotated with their modern-day equivalent, lemma and part-of-speech. The paper presents the structure of the text collection, the sampling procedure, annotation process and encoding of the corpus. The corpus is meant to facilitate HLT research and enable corpus based diachronic studies for historical Slovene. The corpus is encoded according to the Text Encoding Initiative Guidelines (TEI P5), is available via a concordancer and for download from *http://nl.ijs.si/imp/* under the Creative Commons Attribution licence.

Keywords: Reference corpus, Historical language, Slovene.

1. Introduction

Human language technology support for historical language enables diachronic corpus-linguistic studies, better accessibility of cultural heritage texts in digital libraries and better OCR of old books. As opposed to modern language, processing of historical language brings with it a number of problems related to automatic processing:

- due to the low print quality, optical character recognition (OCR) produces much worse results than for modern day texts; currently, such texts must be hand-corrected to arrive at acceptable quality levels;
- full-text search is difficult, as the texts are not lemmatised and use different orthographic conventions with different archaic spellings, typically not familiar to the user;
- comprehension of the texts for most users can also be problematic, esp. with older texts which use different alphabets from the contemporary one.

Diachronic reference corpora typically contain proof-read texts, where each word-form token is annotated with its modern-day equivalent, modern-day lemma and part-of-speech tag. On the basis of such corpora, lexica of historical word-forms can be extracted, and models for spelling change, lemmatization and tagging can be developed. They also serve as the basis for quantitative, corpus based investigations of diachronic language.

Building of annotated corpora of historical language has already been undertaken for a number of languages, e.g., English (Kroch et al., 2004), German (Scheible et al., 2011), Icelandic (Wallenberg et al., 2011) and Spanish (Sánchez-Marco et al., 2010). This paper presents a similar attempt for Slovene, at producing a gold standard and available corpus of historical Slovene, containing facsimiles and word-level annotated transcriptions of sampled pages.

Developing such a corpus, which had so far been lacking for Slovene and annotation tool for historical Slovene is a timely undertaking, as a large number of old books and periodicals are being made available on the Internet, in the context of the Slovene dLib.si digital library, Google books and projects such as the "Slovene literary classics" from the University of Ljubljana making proof-read texts available on WikiSource.

The rest of the paper is structured as follows. Section 2 explains the construction of the corpus, Section 3 its annotation and Section 4 its encoding. Section 5 details the availability of the corpus, gives some conclusions and directions for further work.

2. Corpus construction

2.1. Text collection

The basis for the reference corpus came from a large text collections which comprised proof-read texts with facsimiles:

- Successive selected pages from three religious books, from the end of the 16th, 17th and 18th centuries respectively. The scans of the books and proof-read transcriptions were provided by the Scientific Research Center of the Slovenian Academy of Sciences and Arts. The transcription was initially in Word and then semi-automatically converted to TEI P5. The first two of these books also represent the oldest material in the corpus, barely comprehensible to today's speakers.
- 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, the National and University Library of Slovenia. The books were encoded in PAGE (Pletschacher and Antonacopoulos, 2010), a format 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.
- Selected issues of one Slovenian newspaper, first published in 1843, and continuing to 1890. The facsimiles and transcriptions were also provided by NUK in PAGE.

- The AHLib digital library (Prunč, 2007), 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. 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.

2.2. Sampling

The corpus consists of individual pages sampled from the text collection. Sampling by page rather than by some linguistic unit, as is typically the case for modern-day corpora, was done for several reasons. Each page of the corpus comes with its associated facsimile, and it was simpler to have a one-to-one mapping between the sample unit and facsimile page. The materials that we received from NUK were also individual pages, and due to the design of the PAGE format, it is often difficult to correctly merge individual pages into connected texts; furthermore, at the time of the corpus construction, we did not yet have the complete data-set from NUK. Finally, and perhaps most importantly, it was not obvious which linguistic unit we could have chosen as the basis of the sampling. Chapters are too long, as they will typically span a number of pages and are also very unequal in length between publications. Sentences are very short, and well as being automatically determined, so there is a considerable number of errors in sentence segmentation, also making them unsuitable as a unit for sampling. The most obvious choice would have been paragraphs, as was in fact done in our corpus of contemporary Slovene (Erjavec et al., 2010a). However, many older books do not divide text into paragraphs, or the paragraphs are very long; and in certain text types, such as plays, they are, conversely, very short, making them also unsuitable as the unit of sampling. Of course, splitting the text into page brings with it the disadvantage that a page can, and typically does, start and end in the middle of a paragraph, sentence or even word, compromising the integrity of the linguistic units. To alleviate these problems we have marked potentially split paragraphs and sentences with an attribute (c.f. next section), while split words are annotated with a special tag.

The page sampling procedure tried to ensure, given the sizes and composition of the text collection, a balanced representation of texts and genres. Thresholds were set to limit the maximum number of pages of each particular text, as well as for each decade. Each page also had to contain a minimal number of words, to avoid text-poor pages. With these parameters, the pages were then shuffled, and from this random sequence pages were taken in order, until the desired number of pages (1,000) was collected. Overall, 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 corpus according to the time period, and overall, by the number of units (book or yearly collection of the newspaper), the number of pages (the unit of sampling), and the approximate number of tokens (words or punctuation). The size of the corpus was set to 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	6,000
1695	1	27	10,000
1751-1800	8	155	27,000
1801-1850	12	206	74,000
1851-1875	36	380	126,000
1876-1900	23	224	51,000
Σ	81	1,000	296,000

Table 1: Corpus size by time period

3. Corpus annotation

3.1. Automatic annotation

The corpus was first automatically annotated, using the ToTrTaLe tool (Erjavec, 2011), which tokenises the text, sentence segments it, transcribes historical words to their contemporary form, and tags them with morphosyntactic descriptions and lemmas. For tagging and lemmatization the tool uses models trained on contemporary Slovene, so the transcription step is not only useful by itself, but also crucial for these two levels of annotaiton. The transcriptions is operationalised by the Vaam (Variant Approximate Matching) finite-state library (Reffle, 2011) 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 then, with the help of the LeXtractor editor (Gotscharek et al., 2009) assigned contemporary word-forms to the most frequent (and, typically, unpredictable) words in the collection (Erjavec et al., 2010b). With this static lexicon and transcription patterns we then automatically annotated the reference corpus.

3.2. Manual annotation

In the second step the automatically assigned annotations were manually checked and corrected. The annotation editor used was CoBaLT (Kenter et al., 2012), a Web based corpus browser / editor, in which it is possible to load pre-annotated corpora, correct the annotations as well as the transcriptions, and do this in a concordance-oriented view, so all the occurrences of the same word-form can be inspected and annotated together. A team of annotators, most of them students involved in previous annotation projects, were hired, while the oldest three books were annotated by PhD students in historical Slovenian. The CoBaLT 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. To help with the annotation, the latest hand-corrected corpus was also 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 concordancer has a dedicated front-end, while CQP (Christ, 1994) is used as the back-end.

The corpus was annotated with a view to extracting a lexicon from it, 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 equivalent(s); on the computational lexicon side, historical/modern word boundaries are carefully brought into correspondence (tokenization), abbreviations (sentence segmentation) and foreign passages (tagging and lemmatization) were identified, as were typos in the source. The manual annotation thus corrected mistakes in the transcriptions and tokenization, the contemporary word-form equivalents and lemmas, and added glosses to extinct words.

The tagging was also corrected, but the full morphosyntactic tagset for Slovene contains 1900 different tags (Erjavec et al., 2010a), and is therefore complicated to master and apply. As the work was focused on transcription and lemmatization, we reduced the tagset used in the corpus to a coarse-grained one, which retains only (some) lexical features and has only 33 different tags. Of these, we mention only the tags for the Residual category (X) here, where we distinguish Xf for foreign words, Xt for typos in the source facsimile, and Xp for "program" errors, in particular the parts of the words at the beginning and start of pages.

4. Corpus encoding

The corpus consists of facsimiles, which are of varying quality, but all good enough for on-line reading. The facsimiles not only provide the "base reality" of the texts in the corpus but are also fascinating in their own right, as they contain interesting typefaces, ornaments, and illustrations. Each facsimile is statically stored on the web server in its original format, as well as in two smaller sizes, one for on-line viewing, the other as a thumbnail.

The corpus is encoded as a TEI P5 document (TEI Consortium, 2007), giving the meta-data of the corpus, and links to the 1,000 files corresponding to the transcriptions of individual pages. Figure 1 gives an example of the encoding for (parts of) one page. The page is encoded as a `div` element, where the attributes specify the TEI namespace, the type of the division ("page break"), its language / script identifier (in this case, Slovene written in the "Bohoričica" alphabet, used before 1850) and the link to the facsimile.

The file then gives basic bibliographical information about the work the page is sampled from, followed by a series of `ab` (anonymous block) elements. The `@type` attribute on `ab` specifies what kind of block this is (heading, paragraph, list item, note, etc.). The reason for not using equivalent TEI elements directly is that TEI expects structured documents (e.g., a heading can appear only at the start of a division), and given that the corpus is organised per-page, such structural markup is missing. The `ab` elements then have an identifier attribute, as well as the attribute `@part`, specifying this is potentially only a part of a paragraph, either final (at the top of the page) or initial (at the bottom). Furthermore, the `@corresp` attribute on `ab` gives a pointer into the PAGE file, where the coordinates of the facsimile region in which this block of text appears are given.

The blocks are then marked-up with linguistic information, i.e. sentences, and these of words, punctuation symbols, and whitespace. Words have attributes for the normalised form (`@nform`), modernised form (`@mform`), lemma (`@lemma`), and corpus tag (`@ctag`). Where the word does not exist anymore in the contemporary language only its spelling is modernised. However, the closest contemporary synonyms are given in the `@gloss` attribute.

```
<div xmlns="http://www.tei-c.org/ns/1.0"
     type="pb" xml:lang="sl-boh"
     xml:id="goo18B-NUKR10214-1790.pb.095"
     facs="facs/NUKR10214-1790/00422752_m.jpg">
   <bibl>
      <title>Shupanova Mizka</title>
      <author>Linhart, A. T.</author>
      <date>1790</date>
   </bibl>
   <ab type="p" part="F"
       corresp="NUKR10214-1790/00422752.xml#r6"
       xml:id="goo18B-NUKR10214-1790.ab.1067">
      <s>
         <w nform="baron" mform="baron"
            lemma="baron" ctag="Ncm">Baron</w>
         <pc ctag=".">.</pc>
      </s>
      <c> </c>
      <s>
         <w nform="vſmili" mform="usmili"
            lemma="usmiliti" ctag="Vme">Vſmili</w>
         <c> </c>
         <w nform="ſe" mform="se" lemma="se"
            ctag="P">ſe</w>
         <pc ctag="!">!</pc>
      </s>
      ...
      <s>
        <w nform="shentani" mform="šentani"
           lemma="šentan" ctag="Agp">Shẹntani</w>
        <c> </c>
        <lb n="6"/>
        <w nform="keklavez" mform="kekljavec"
           ctag="Ncm" lemma="kekljavec"
           gloss="jecljavec">keklavez</w>
        <c> </c>
        <pc ctag="!">!</pc>
      </s>
      ...
   </ab>
</div>
```

Figure 1: Example of the TEI encoding of a corpus element.

```
<w nform="ma" mform="midva" lemma="midva"
  ctag="P" n="mw_694">ma</w>
<c> </c>
<w nform="dua" mform="midva" lemma="midva"
  ctag="P" n="mw_694">dua</w>

<w nform="nevtikui" mform="ne_vtikuj"
   lemma="ne_vtikovati"
   ctag="Q_Vmp">nevtikui</w>
```

Figure 2: Encoding of joined and split words.

A special case arises with joined or split words, i.e. situations where several historical words correspond to one contemporary one or vice versa, as illustrated in Figure 2. The former case is encoded with two (or more) word tokens all having the same modern form, lemma and corpus tag, but

being associated via the value of the `@n` attribute. The latter case is modeled as one word token, where the modern form, lemma and corpus tag have portmanteau values.

5. Conclusions

The paper presented the goo300k linguistically annotated corpus of historical Slovene, giving its composition, annotation and encoding. The corpus and supporting documentation are available from `http://nl.ijs.si/imp/` for concordancing, as well as for download under the Creative Commons, Attribution Licence. The liberal licence is made possible by the fact that the texts are out of copyright, while the producers of the proof-read transcriptions and structural annotations (the Austrian Academy of Sciences and the National and University Library) have kindly agreed to make them available under the CC-BY licence.

We hope that this corpus will provide a catalyst for corpus-based linguist studies and for research on computational processing of historical Slovene. This would enable digital libraries to develop better information retrieval for Slovene cultural heritage texts and developers of OCR software to better capture them.

Further work includes enlarging the corpus and extracting a reference lexicon from the corpus, encoding it in TEI and putting it on-line as a browsable and searchable resource. We also plan to re-train ToTrTaLe on the corpus and lexicon, automatically annotate the complete text collection and make this available for search and download as well. More research oriented work includes concentrating on the challenging aspects of the modernisation procedure, i.e. tokenizaton mapping, automatically inducing transcription rules, as well as "translating" the corpus into contemporary Slovene and training statistical machine translation models to encompass syntactic changes between historical and contemporary language.

Acknowledgments

The author would like to thank the anonymous reviewers for their helpful comments and suggestions. The work presented in this paper has been supported by the EU IMPACT project "Improving Access to Text" and the Google Digital Humanities Research Award "Developing language models for historical Slovene".

6. References

Oliver Christ. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of COMPLEX '94: 3rd conference on Computational Lexicography and Text Research*, pages 23–32, Budapest, Hungary.

Tomaž Erjavec, Darja Fišer, Simon Krek, and Nina Ledinek. 2010a. The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Tomaž Erjavec, Christoph Ringlstetter, Maja Žorga, and Annette Gotscharek. 2010b. Towards a Lexicon of XIXth Century Slovene. In *Proceedings of the Seventh Language Technologies Conference*, Ljubljana, Slovenia, October. Jožef Stefan Institute.

Tomaž Erjavec. 2011. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 33–38, Portland, OR, USA, June. Association for Computational Linguistics.

Annette Gotscharek, Andreas Neumann, Ulrich Reffle, Christoph Ringlstetter, and Klaus Schulz. 2009. Enabling Information Retrieval on Historical Document Collections — the Role of Matching Procedures and Special Lexica. In *Proceedings of the ACM SIGIR 2009 Workshop on Analytics for Noisy Unstructured Text Data (AND09)*, Barcelona.

Tom Kenter, Tomaž Erjavec, Maja Žorga, and Darja Fišer. 2012. Lexicon construction and corpus annotation of historical language with the CoBaLT editor. In *Proceedings of the EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Avignon, France, April. Association for Computational Linguistics.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/histcorpora/PPCEME-RELEASE-2/.

Stefan Pletschacher and Apostolos Antonacopoulos. 2010. The PAGE (Page Analysis and Ground-truth Elements) Format Framework. In *Proc. of the 20th International Conference on Pattern Recognition (ICPR)*. Istambul.

Erich Prunč. 2007. Deutsch-slowenische/kroatische Übersetzung 1848-1918. Ein Werkstättenbericht. (German-Slovene/Croatian translation, 1848-1918. A workshop report). *Wiener Slavistisches Jahrbuch*, (53):63–176.

Ulrich Reffle. 2011. Efficiently generating correction suggestions for garbled tokens of historical language. *Nat. Lang. Eng.*, 17:265–282.

Cristina Sánchez-Marco, Gemma Boleda, Josep Maria Fontana, and Judith Domingo. 2010. Annotation and Representation of a Diachronic Corpus of Spanish. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. A Gold Standard Corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 124–128, Portland, Oregon, USA, June. Association for Computational Linguistics.

TEI Consortium, editor. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*.

Joel Wallenberg, Anton Karl Ingason, Einar Freyr Sigurthsson, and Eiríkur Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC), Version 0.9. http://www.linguist.is/icelandic_treebank.