Language Resources and Evaluation
DOI 10.1007/s10579-015-9294-7

The IMP historical Slovene language resources

Tomaž Erjavec

Published online: 21 January 2015
http://dx.doi.org/10.1007/s10579-015-9294-7

Abstract The paper describes the combined results of several projects which constitute a basic language resource infrastructure for printed historical Slovene. The IMP language resources consist of a digital library, an annotated corpus and a lexicon, which are interlinked and uniformly encoded following the Text Encoding Initiative Guidelines. The library holds about 650 units (mostly complete books) consisting of facsimiles with 45,000 pages as well as hand-corrected and structured transcriptions. The hand-annotated corpus has 300,000 tokens, where each word is tagged with its modernised word form, lemma, part-of-speech and, in cases of archaic words, its nearest contemporary equivalents. This information was extracted into the lexicon, which also covers an extended target-annotated corpus, resulting in 20,000 lemmas (of these 4,000 archaic) with 50,000 modern word forms and 70,000 attested forms. We have also developed a program to modernise, tag and lemmatise historical Slovene, and annotated the digital library with it, producing an automatically annotated corpus of 15 million words. To serve the humanities, the digital library and lexicon are available for reading and browsing on the web and the corpora via a concordancer. For language technology research and development the resources are available in source TEI XML under the Creative Commons Attribution licence. The paper presents the IMP resources, available from http://nl.ijs.si/imp/, the process of their compilation, encoding and dissemination, and concludes with directions for future research.

Keywords Historical language resources · Slovene language · Text Encoding Initiative · Non-standard language normalisation

Tomaž Erjavec
Dept. of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana
Slovenia
E-mail: tomaz.erjavec@ijs.si

1 Introduction

With the rapidly increasing volume of digitally available written cultural heritage, recent years have seen an explosion of research on and development of corpora and processing tools for a number of historical languages and with a variety of methods (Piotrowski, 2012). Reference corpora and annotation tools were developed first for English (Kroch et al., 2004; Rayson et al., 2007) and major European languages, e.g., German (Scheible et al., 2011) and Spanish (Sánchez-Marco et al., 2010), as well as for some smaller ones, in particular Icelandic (Wallenberg et al., 2011). Some resources for Slavic languages, distinguished by their rich inflectional morphology, have also been produced, in particular for Czech (Kučera, 1999) and and Polish (Dudczak et a.l, 2012; Bień, 2014). This paper presents a set of historical resources for Slovene, a language for which their availability has been quite limited so far. The resources, called IMP, consist of a large digital library, also available as an automatically annotated corpus; a computational lexicon, also deployed as an on-line dictionary; a manually annotated corpus; and an annotation pipeline and support tools for corpus compilation.

The IMP resources span the period 1584–1918, however, the bulk of the texts is relatively recent, from 1850 onwards. This comparatively young language is nevertheless interesting as it contains many Slovene classics as well as much more varied text types than earlier periods, where the vast majority of texts were of a religious nature. Furthermore, Slovene spelling was standardised to today's norm quite recently. Even texts from the second half of the 19^{th} century contain over 10% of word tokens written differently than today, so simply using processing tools for contemporary language gives low quality results. There is also orders of magnitude more material available here than for earlier periods, so developed methods for text processing have much more scope for application.

The IMP resources are distinguished by the ambition to be useful for both human language technologies (HLT) as well as digital humanities (DH). For DH the resources have been converted to various presentation formats and mounted on the Web where they are available for reading, searching and corpus exploration. For HLT we aimed to create a harmonised "basic language resources toolkit" or BLARK (Krauwer, 2003) for historical Slovene. To this end the resources are available for download in their source XML under the very permissive Creative Commons Attribution (CC BY) licence, so they can be used not only for research but also for developing open source or commercial applications, such as improving OCR of old Slovene texts, supporting better information retrieval in cultural heritage digital libraries, and modernising old texts to make them better understandable to today's readers. Apart from offering a large collection of proof-read texts accompanied by facsimiles, the focus of the HLT effort was on orthographically modernising individual words in the texts, and, in cases of archaic words or those exhibiting semantic shift, giving them glosses or synonyms in the contemporary standard language.

The rest of this paper is structured as follows: Section 2 gives a short history of the Slovene language; Section 3 introduces the TEI encoding of the resources; Section 4 details the collection of historical Slovene texts available as a digital library, gives the sources of these texts, which mostly correspond to several related projects

that enabled the work on IMP, and presents their up-translation to TEI; Section 5 concentrates on the manually annotated corpus, explaining the annotation work-flow and annotation types; Section 6 presents the corpus-based lexicon; Section 7 briefly discusses the tool developed to annotate old Slovene and the automatically annotated corpus produced from the digital library; and Section 8 gives some conclusions and directions for further research.

2 A short history of Slovene

Slovene (or Slovenian), a South Slavic language, is the official language in Slovenia, with about 2 million speakers. It is distinguished by its rich inflection, e.g., it distinguishes six cases, three genders and three numbers (still retaining the dual), and exhibits a complex system of inflectional paradigms with many subregularities and exceptions.

The first Slovene books were published in the second half of the 16^{th} century, including the translation of the Bible and the first grammar book. Up to 1918, the territory of present-day Slovenia was a part of the Austro-Hungarian Empire, with German being the language of the elite, and Slovene the vernacular. German thus had a strong influence on Slovene, however, the rise of Romantic Nationalism in the 19^{th} century brought words from Serbo-Croatian and Czech into standard Slovene, mostly to replace words previously borrowed from German. At the end of the 19^{th} and beginning of the 20^{th} century, in the Fin de siècle period, the first generation of Slovene modernist authors reversed this tendency by using a more "pure" Slovene language without many Serbo-Croatian borrowings.

Until the mid-19^{th} century Slovene used the Bohorič alphabet, which was modelled on the German one, using *ſ* for today's *s*, *z* for *c*, *s* for *z*, *zh* for *č*, *ſh* for *š*, and *sh* for *ž*. All the letters of the old alphabet, except *ſ*, are still used today but they correspond to different sounds, which makes reading texts in the Bohorič alphabet difficult for today's readers.

Already from the beginning of the 19^{th} century various Slovenes argued for a new alphabet, more in line with the phonological structure of Slovene and taking alphabets of other Slavic languages as the model. In the 1830s Slovenia entered a "spelling war", with various new alphabets, in particular the Dajnko and Metelko alphabets, trying to supplant the Bohorič one. The alphabet that eventually won and was used almost exclusively after a relatively brief transition period in 1843 is the Gaj alphabet. It is the one still used today, and employs standard Latin letters (*a-z* except for *q, w, x, y*) and *č*, *š* and *ž*.[1]

To illustrate how the language was changing, and how it differs from contemporary Slovene, Figure 1 gives three text snippets written roughly 50 years apart; all examples contain the word *ljubezen* (*love*).

As seen in the examples the spelling changes in Slovene have been considerable, even apart from the alphabet change; the first example has three (out of 11) words

[1] We have registered with IANA the `sl-bohoric` sub-language tag for texts using the Bohorič alphabet, as well as `sl-dajnko` and `sl-metelko` for the Dajnko and Metelko alphabets, in which a number of books were also printed. These tags can be used e.g., as the value of `@xml:lang` attribute.

1790	Al	ta	nar bòl	vashna	reſsniza	je	moja	lubęsen	prut	Nęshki.
(18B)	*ali*	*ta*	*najbolj*	*važna*	*resnica*	*je*	*moja*	*ljubezen*	*proti*	*nežki*

1843	poboshnim	ſerzam	in	veſtjo	pridnoſt	in	ljubesin	k	ſvojimu	ſtanu	sdrushi
(19A)	*pobožnim*	*srcem*	*in*	*vestjo*	*pridnost*	*in*	*ljubezen*	*k*	*svojemu*	*stanu*	*združi*

1872	Otroška	ljubezen	naj	zmír	te	navdaja	Za	starše,	za	brate,	Bogá	in	cesarja
(19B)	*otroška*	*ljubezen*	*naj*	*zmeraj*	*te*	*navdaja*	*za*	*starše,*	*za*	*brate,*	*boga*	*in*	*cesarja*

Figure 1 Slovene text from different periods, with word modernisation.

spelled differently from the contemporary norm, the second two, and the third one. Furthermore, due to the lack of standardisation and dictionaries, authors from the same period often used different spellings, and even texts from the second half of the 19^{th} century contain, on average, over 10% of word tokens with archaic spellings. To give some more insight into the variability of historical Slovene spelling and of the inflectional nature of Slovene, we take the word (lemma) *ljubezen* (*love*): in the IMP corpus there are 21 different spellings of this particular word, 5 attested (contemporary) inflectional word forms and 42 spellings of these word forms.

A not uncommon phenomenon, exemplified by the modern word form *najbolj* in the first example of Figure 1, is that two (or more) word forms that used to be written apart are now written as one word, or vice versa. As will be discussed, this presents a problem for the encoding and annotation focused on modernising individual words. Apart from spelling changes, many historical words, such as those imported from German or Serbo-Croatian, are no longer used, various words have exhibited a significant semantic shift or change in their morphosyntactic properties, from gender change to different inflectional endings; such words can be modernised in terms of spelling, but they need an added gloss to make them either understandable to today's speaker or to function as replacement words in searches.

3 Resource encoding

The complete set of IMP resources is encoded following the Text Encoding Initiative Guidelines TEI P5 (TEI Consortium, 2012), a parametrisable annotation vocabulary and encoding scheme, with which it is possible to construct project-specific XML schemas and use them to validate and document the resource encoding. While many other standards (in particular those developed by ISO TC37 SC4) and de-facto encoding schemes exist, none of them would cover all the varied text types of the IMP resources, so we would otherwise need to use a host of disparate standards to encode them all.

To illustrate the structure of a TEI document, Figure 2 gives the skeleton of the oldest book in the digital library, with its structure similar to those of the other IMP resources. TEI documents have a book-like structure, with the root `TEI` element giving the TEI namespace and, in our case, the identifier of a particular text. The first

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="ZRC_00001-1584">
   <teiHeader xml:lang="sl"> ... </teiHeader>
   <facsimile> ... </facsimile>
   <text xml:lang="sl-bohoric">
     ...
   </text>
</TEI>
```

Figure 2 The basic structure of a TEI document.

contained element is always the `teiHeader` which gives the text meta-data and also specifies that the (main) language of the meta-data is Slovene. The meta-data in TEI can be quite extensive, giving the file description, including the description of its source; an encoding description with the practices observed in encoding the document; a profile description including information on e.g., type of the text according to a specified taxonomy; and a revision description explaining the changes to the document. For the texts of the digital library and the corpus, the second `TEI` element is the (optional) `facsimile` giving the locations and properties of facsimile image files followed by the (obligatory) `text` element, which is in Figure 2 in the Bohorič alphabet. In IMP this element contains, depending on the particular resource, an annotated transcription of a text (as is the case in the presented example); the complete lexicon; a text from a linguistically annotated corpus; or the morphosyntactic specifications defining the part-of-speech tagset used in the corpus and lexicon.

By using 11 TEI modules (e.g., linking, figures and tables, transcription, corpora, feature-structures, dictionaries), we made a single XML schema for all the resources: not counting the elements for the `teiHeader` (i.e., meta-data) the digital library uses 45 different elements, the lexicon uses 18, the morphosyntactic specifications 28, the reference annotated corpus 19, and the automatically annotated corpus made from the digital library 42. The complete set of resources uses 75 different TEI elements, each with a well-documented semantics and, in many cases, ready-made XSLT down-conversion into HTML and other formats.

It should be noted that, unlike most other approaches to encoding linguistically analysed texts, esp. automatically annotated ones, the TEI (linguistic) encoding we use is in-line rather than stand-off: all the annotation is on the text instead of pointing to it. While this brings with it the well-known problem of crossing hierarchies, TEI has various methods to deal with it. But, on the other hand, all problems that are latent in a stand-off approach are here immediately identifiable. For example, manual linguistic annotation of historical sources often reveals remaining mistakes in the source transcription, which are very difficult to correct in a stand-off approach without compromising the integrity of the encoding but is much less problematic in-line.

4 The digital library

While there already exist extensive digital libraries of historical Slovene texts, in particular in the scope of Google Books and the main Slovenian digital library dLib.si,

these texts are available only as PDF facsimiles and low-quality OCR transcriptions. In this section we present the IMP digital library (DL), currently the largest digital library of historical Slovene, containing, in addition to facsimiles, carefully proof-read and structurally annotated texts.

The IMP DL documents have extensive meta-data, including the alphabet used, year (span) of the publication, title in the original and in contemporary Slovene, and taxonomic information for the type of text (fiction, non-fiction, religious), text medium (book, magazine, newspaper, manuscript) and translation status (original or translated into Slovene). Each text is encoded as one TEI document, with the text containing structural mark-up, including sections, headings, paragraphs, footnotes, etc. and page breaks with links to associated page scans, which are available in several sizes.

The TEI P5 library has been converted to high-quality HTML with slightly customised TEI XSLT stylesheets[2] and mounted on the web as a collection of static files (one per text) with several indexes. The TEI headers containing the meta-data about each text have also been converted to HTML, with the names of elements glossed to Slovene.[3] The Web library has anchors at individual pages, so it is possible to refer with stable URLs to specific pages in the publications.

An overview of the size of the IMP DL along various dimensions is given in Table 1. The overall size is 658 units amounting to almost 45,000 pages or over 14 million words. The distribution over time of publication shows that the oldest units are from the end of the 16^{th} and end of the 17^{th} centuries, but these are only small samples from two books. In retrospect, including these samples might have been a mistake, as they were very difficult to annotate with linguistic information (explained further in Section 5) while not offering enough data to build good models or, indeed, be useful for reading. More materials are available from 1768–1849, with 41 units (mostly complete books) and over 1.2 million words. While this might not seem a large amount, it has to be considered that Slovene did not have much text production at the time. The bulk of the materials (90%) is from 1850–1918 and contains the most varied texts, both in terms of text type and medium.

The size of the DL by taxonomy is given in the next three sections of the table. Half the words in the library come from books, and over 40% from magazines, which was a popular and very common way of publishing both non-fiction as well as fiction by Slovene authors, esp. from the 2^{nd} half of the 19^{th} century onwards. The DL also contains 63 samples of one newspaper, containing over 1 million words (cf. Sec. 4.3), as well as a 11 manuscripts, one from 1776 and the others from 1854 and 1855.

The DL has 32 religious texts with just under 1 million words, and these constitute most of the older materials in the library. About one fifth or 3 million words is non-fiction of various genres, (cf. Sec. 4.1), while the majority is fiction (75% of the texts or 70% of the words), further subdivided into some plays and text containing predominantly poetry, with all the rest being prose. Finally, the texts are also divided into those that were authored in Slovene, and into translated works, which

[2] TEI stylesheets are available at `http://www.tei-c.org/Tools/Stylesheets/` and support the conversion of many formats to and from TEI P5.

[3] We have given Slovene glosses to all the 546 elements defined by the TEI. The localisation is available at `http://nl.ijs.si/tei/locale/`.

	Texts		Pages		Words	
	658	100.0%	44,951	100.0%	14,473,005	100.0%
1584	1	0.2%	8	0.0%	5,880	0.0%
1695	1	0.2%	27	0.1%	8,619	0.1%
1768–1799	14	2.1%	2,345	5.2%	379,011	2.6%
1800–1849	27	4.1%	3,653	8.1%	887,589	6.1%
1850–1899	368	55.9%	22,647	50.4%	7,400,493	51.1%
1900–1918	247	37.5%	16,271	36.2%	5,753,403	39.8%
book	275	41.8%	29,157	64.9%	7,084,733	49.0%
magazine	309	47.0%	13,850	30.8%	6,085,893	42.0%
newspaper	63	9.6%	1,674	3.7%	1,175,004	8.1%
manuscript	11	1.7%	270	0.6%	89,365	0.6%
religious	32	4.9%	4,889	10.9%	988,694	6.8%
nonfiction	127	19.3%	8,661	19.3%	3,004,279	20.8%
play	21	3.2%	2,379	5.3%	360,153	2.5%
poetry	3	0.5%	906	2.0%	191,138	1.3%
prose	475	72.2%	28,116	62.5%	9,890,731	68.3%
original	547	83.1%	32,316	71.9%	11,583,141	80.0%
translation	111	16.9%	12,635	28.1%	2,851,854	19.7%

Table 1 Composition of the IMP library.

constitute almost 17% of the texts or 20% of all the words in the DL. This relatively large amount of translated works is very welcome, as, on the one hand, cultural heritage libraries usually concentrate on natively authored texts, while, on the other hand, Slovenes have always read a lot of translated texts, so including them makes for a more balanced resource collection, as well as being important for (diachronic) translation studies (cf. Sec. 4.1).

The texts for the digital library came from four sources, which roughly correspond to four different but connected projects involving the Jožef Stefan Institute (JSI), and which also have quite different characteristics in terms of text type and age, leading to a much more varied offering than could have been expected otherwise. However, they also made producing the IMP DL more difficult, as the texts from the various sources were produced in different ways, so their up-translation to TEI P5, including the kinds of markup that are included in the text, had to be performed differently. These sources, the kinds of materials they contain, their conversion to TEI and the specifics of their mark-up are discussed next.

4.1 The AHLib digital library

AHLib (Austro-Hungarian digital Library) was one of the results of the project "German to Slovene translation in the period 1848–1918" led by Prof. Erich Prunč at the Austrian Academy of Sciences and Graz University (Prunč, 2007). AHLib was the first attempt at developing language resources for historical Slovene (Erjavec, 2007), and the complete AHLib DL is included in IMP.[4]

This sub-library contains 105 books translated from German and chosen as the most historically or linguistically important works from this period. Most AHLib

[4] The site with the AHLib DL and links to the Graz resources is `http://nl.ijs.si/ahlib/`.

books are fiction, often with a strong educational and moralistic component, as was popular in Austro-Hungary. Probably the best known here are the books by C. von Schmid, containing e.g., *Zgodbe Svetega pisma za mlade ljudi* (*Stories from the Bible for young people*) (1830, 196pp); also noteworthy is the book *Robinson mlajši* (*Robinson the Younger*) (1849, 367pp), known as the best exemplar of philanthropist pedagogy, and with a highly idiosyncratic translation, which includes a very large number of Croatian words and uncommon spellings, significantly differing from the norms of the time.

In terms of the number of pages, almost half of AHLib is non-fiction. There are many "how-to" books, e.g., F. Haberlandt, *Kratka navodila za rejo domačih svilo-prejk* (*Short instructions for raising silkworms*) (1871, 16pp) or A. Trientl, *Stelja in gnoj* (*Litter and manure*) (1875, 23pp). The most interesting non-fiction materials are a number of comprehensive textbooks on natural sciences, in particular books by F. K. Ludwig Schödler, *Astronomija* (*Astronomy*) (1869, 112pp), *Kemija* (*Chemistry*) (1869, 196pp), *Fizika* (*Physics*) (1869, 230pp), etc. The books have many illustrations and are a good source for diachronic terminology studies.

The AHLib texts were first proof-read and type-set in Word and then converted to TEI via a dedicated rtf2tei Web application.[5] The preparation of the books was supported by several training sessions and an annotators' manual and was performed by several students engaged on a long-term basis. The quality of the AHLib books is therefore quite high, with few typos and structural mistakes. Any mistakes found in the subsequent HTML rendering of the complete AHLib were also corrected.

In addition to structural markup the books also contain basic linguistic and editorial annotations, such as text passages in a foreign language and corrections to typos. The position, captions and descriptions of illustrations are also included in the transcriptions, which enables simple identification of page scans containing figures or page ornaments. To illustrate these points and the general encoding of the DL we give in Figure 3 an example from the TEI encoding of an AHLib book. The `body` (i.e., text proper) of the books starts with the page break containing a pointer to the facsimile (`pb/@facs`). As was often the case at the time, this front page contains an illustration followed by its caption; while the exact positioning of the illustrations has not been marked-up in AHLib, the fact that a certain page contains illustration(s) is annotated, together with a description of the illustration. The second page then starts with the text, which is structured into divisions, headings, paragraphs, etc.

4.2 The SRC-SASA samples

In the scope of the Google Digital Humanities Research Award given jointly to the Scientific and Research Centre of the Slovene Academy of Sciences and Arts (SRC-SASA) and to JSI, SRC-SASA prepared samples of three religious books, 75 pages

[5] While the original rtf2tei converter is no longer maintained, we have developed a new web-based Word to TEI P5 converter, which upgrades the current TEI XSLT stylesheets for docx2tei and tei2html conversion. The Web service accepts Office Open XML documents and converts them to TEI P5 and from there to HTML, and stores the complete results on a unique URL. It is available on `http://nl.ijs.si/tei/convert/`.

```
<body>
  <pb n="[0]" facs="#FPG00008-000" xml:id="pb.001"/>
     <figure xml:id="figure.1">
        <figDesc>Svetnik s kelihom v levi roki, na kelihu kača,
                 desno roko ima rahlo privzdignjeno</figDesc>
      </figure>
      <p xml:id="p.1">BRATERNA SVETIGA JANŠA.</p>
      <pb n="[1]" facs="#FPG00008-001" xml:id="pb.002"/>
      <div type="level1" xml:id="div.1">
        <head type="docTitle" xml:id="head.1">ČUJTE, ČUJTE
             KAJ ŽGANJE DELA!</head>
        <p xml:id="p.2">Prigodba žalostna ino vesela za
             Slovence.</p>
        <p xml:id="p.3">Poslovenil</p>
        <p xml:id="p.4">F... GL...</p>
```

Figure 3 Example from a TEI encoded AHLib book.

in total. Two of these books, namely J. Dalmatin, *Biblija* (*The Bible*) (1584, 8pp) and J. Svetokriški, *Sacrum promptuarium* (1695, 27pp), are also the oldest works in IMP.

The preparation of these samples proceeded via Word, in the same way as for AHLib; due to the small size of the texts and highly qualified proof-readers (two postgraduate students in the history of Slovene language) the texts are most likely error-free.

4.3 The National Library Ground Truth Data

In the scope of the EU IMPACT project the National and University Library of Slovenia (NUL) prepared 5,000 pages of proof-read texts (Erjavec et al., 2011). Of these, 14 are AHLib books which had been scanned but not transcribed in the scope of the original project, while 12 are some of the most important Slovene books from the late 18^{th} and early 19^{th} centuries, for example A. T. Linhart, *Županova Micka* (*Mayor's Mary*) and *Ta veseli dan ali Matiček se ženi* (*This happy day or Matiček gets married*) (1790, 187pp), the most famous Slovene plays from the 18^{th} century, or V. Vodnik, *Kuharske bukve* (*Books on cooking*) (1799, 207pp), the first Slovene cookbook. The NUL collection also contains 590 pages that were sampled from 47 years (1843–1890) of the newspaper *Kmetijske in rokodelske novice* (*Agricultural and Artisan News*).

The OCR of these texts was corrected and edited in the Aletheia editor (Clausner et al., 2011), meant for producing ground-truth data for OCR-system development. The annotators and proof-readers were regular employees of NUL, who also performed several internal checks and evaluations on the quality of their work, and estimate the accuracy of the transcription at 99.95% (Erjavec et al., 2011).

The output format of Aletheia follows PAGE XML (Pletschacher and Antonacopoulos, 2010), which is a customisation of ALTO (Analyzed Layout and Text Object), a page oriented annotation scheme supporting mark-up of scan polygons and associating them with blocks of text. These blocks can be ordered and assigned labels such

```
<facsimile n="20">
  <graphic n="orig" url="http:.../NUKP14041-1851.pdf"/>
  <surface xml:id="GTD-00423318"
              ulx="4769" uly="6564" lrx="0" lry="0">
    <desc>NUKP14041-1851 [1]</desc>
    <graphic n="orig" url="http:.../00423318.tif"
                     width="4769px" height="6564px"/>
    <graphic n="medium" url="http:.../00423318_m.jpg"/>
    <graphic n="small" url="http:.../00423318_s.jpg"/>
    <graphic n="thumb" url="http:.../00423318_t.jpg"/>
    <zone xml:id="GTD-00423318.r3"
         points="4644,111 4644,839 ... 222,111"/>
 ...
       <body>
         <pb xml:id="pb.001" n="1" facs="#GTD-00423318"/>
         <div xml:id="div.1">
            <head facs="#GTD-00423318.r3">NOVICE</head>
            <head facs="#GTD-00423318.r129">kmetijskih, rokodelnih in
              narodskih rečí.</head>
```

Figure 4 Example of facsimile encoding in a NUL text.

as heading, paragraph, footnote, printer's mark, etc. We developed an XSLT pipeline to convert PAGE XML to TEI P5 but as the source format is not structurally oriented, this component of the corpus has less sophisticated structural markup, e.g., divisions are flat, rather than hierarchical. On the other hand, facsimile-transcription alignment is more fine-grained, as it is not only by page but by paragraph-level elements.

To illustrate, we give in Figure 4 the start of the TEI `facsimile` element, followed by the start of the text `body` for one of the NUL texts. The facsimile first contains the `graphic` giving the URL (shortened for readability) of the PDF containing the complete facsimile of the text. This is followed by the first page, encoded as the `surface` element, which gives, via its attributes, the identifier and scale of the facsimile. Each page has an automatically generated description and pointers to the images in four sizes, in `desc` and `graphic`, respectively. Up to this point, the facsimile and page annotations are the same as for all the other IMP texts. However, NUL has the additional marking of zones on the surface, which are defined by polygons (here shortened). As can be seen in the body of the text, each structural element then refers to the identifier of its zone.

4.4 Wikivir literary classics

By far the largest (and the most recently added) part of IMP comes from the Wikisource ("Wikivir" in Slovene) project "Slovene literary classics"[6] led by Prof. Miran Hladnik from the University of Ljubljana (Hladnik, 2009). For a number of years the most important out-of-copyright Slovene literary texts (published as books, articles from newspapers or magazines, with an occasional manuscript) are being added to

[6] The project can be found at `http://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika`

Wikivir, first as raw OCR, which students of Slovene language then proof-read and type-set. The students' work is supported by a yearly grant from the Slovene Ministry of Culture, and was additionally supported by our Google Award. In this way an extensive and freely available collection of Slovene classics is being compiled, which currently contains mostly 19^{th} and early 20^{th} century works. The quality of proof-reading varies very much as over a hundred students were involved in this effort over the years, and their skill and attention to detail differed significantly. A spot check showed that the majority of the texts are of good quality, with one or two errors per page, although a few works display as many as 10 errors per paragraph. However, the great advantage of Wikisource is that any errors found by anyone can be immediatelly fixed, and we plan to re-import the corrected and extended Wikivir in the future.

From Wikivir we selected all the units that date before 1918 and have facsimiles, or these could be obtained. This includes many books from the classic Slovene novelists, such as I. Cankar, J. Jurčič, J. Kersnik, and I. Tavčar. Additionally, Simon Atelšek, in the scope of his PhD research at SRC SASA, added four books on beekeeping, which is still very popular in Slovenia, with beehives kept even in urban areas. Two of these books are from the end of the 18^{th} century, and two from 1830s, including Peter Dajnko, *Čebelarstvo* (*Beekeeping*) (1831, 245pp), written, of course, in the Dajnko alphabet; but at least until fonts (and Unicode code-points) are available, it is transcribed into the quotidian Gaj alphabet. In total, the IMP Wikivir currently contains over 30,000 pages or precisely 500 texts of 154 Slovene authors.

In the scope of the Google Award, Domen Kermc (who also led the Wikivir correction work in 2012–2013) developed a program which takes a URL of a Wikivir text as input, uploads the text, and converts it into TEI P5.[7] The application is customised to the particular format of our Wikiproject and is able to extract most of the meta-data for each text as well as preserve the sometimes very detailed typesetting of the original, e.g., various types of highlighting, correction of typos, figures, tables, footnotes, stanzas, etc. All of this mark-up is also preserved in the HTML version of the DL.

5 The goo300k corpus

On the basis of (the initial version of) the IMP DL we compiled a hand-annotated corpus meant to serve as a gold-standard dataset for linguistic annotation of historical Slovene, especially for the 19^{th} and 18^{th} centuries. The corpus was used in IMP for extracting the lexicon (discussed in Section 6), but in a wider perspective it is intended for use as a training and testing dataset for the development of automatic annotation of historical Slovene.

[7] The wiki2tei converter is available as a Web service at `http://nl.ijs.si/wiki2tei`, with its PHP source on `https://github.com/domenk/wiki2tei`. It has an interface both in Slovene and English and can also convert the works of other Wikisource languages, but the quality of the output depends on how much their format differs from the Wikivir one.

	Texts		Pages		Words	
	89	100.0%	1,100	100.0%	290,587	100.0%
1584	1	1.1%	8	0.7%	5,794	2.0%
1695	1	1.1%	27	2.5%	8,519	2.9%
1768-1799	8	9.0%	155	14.1%	22,216	7.6%
1800-1849	17	19.1%	280	25.5%	104,428	35.9%
1850-1899	62	69.7%	630	57.3%	148,413	51.1%
book	78	87.6%	994	90.4%	191,411	65.9%
newspaper	11	12.4%	106	9.6%	97,959	33.7%
religious	22	24.7%	326	29.6%	63,926	22.0%
nonfiction	28	31.5%	310	28.2%	146,855	50.5%
play	11	12.4%	145	13.2%	20,308	7.0%
poetry	2	2.2%	38	3.5%	4,875	1.7%
prose	26	29.2%	281	25.5%	53,406	18.4%
original	22	24.7%	283	25.7%	125,980	43.4%
translation	67	75.3%	817	74.3%	163,390	56.2%

Table 2 Composition of the goo300k corpus.

The development of the corpus was supported by the Google Award and is called goo300k.[8] It contains 1,100 page samples[9] with the distribution given in Table 2.

As can be seen from the temporal distribution of the texts, the complete SRC-SASA samples of the oldest texts from the IMP DL are included in the corpus. For the period 1768–1849 the corpus contains almost half of the materials, i.e., this period is, relatively speaking, better represented than in the IMP DL. Nevertheless, the majority of the texts are from the second half of the 19^{th} century. It should be noted that the youngest materials are from 1899, rather than from 1918 as in the DL: as the language gets progressively closer to the contemporary standard, it seemed more worthwhile to concentrate on somewhat older texts. As for the proportions of the various text media, types and translation status, they represent the best approximation of a balanced corpus that could be achieved given the relatively small DL available at the time the corpus was compiled, lacking all Wikivir texts and with some NUL texts not yet ready. Religious works are rather strongly represented, comprising about a quarter of the corpus, but, as mentioned, this was the predominant text type for older works. Non-fiction gives almost half of the corpus, mostly due to the fact that, at the time, there were much more AHLib texts available than others and that quite a few NUL texts are also non-fiction. The fiction part thus comprises only about a quarter of the corpus. Finally, over half of the corpus is translated materials, again due to the preponderance of AHLib materials.

The corpus is structured into texts, each containing meta-data, the facsimiles of the sampled pages and their transcription. The transcription of each page is then composed of paragraph-level blocks typed with values such as paragraph, heading, caption, etc. When this is the case, the first or last block on the page is marked as incomplete. These blocks were then automatically tokenised and sentence segmented

[8] A previous, smaller and less well-annotated version is described in Erjavec (2012b).

[9] Page sampling is unusual as pages do not correspond to linguistically motivated units but in the case of historical texts it is difficult to come up with a better alternative as some texts do not even distinguish divisions or paragraphs or these are very long. Furthermore, the alignment of text samples with the facsimiles is also preserved in this way.

```
<ab type="p" xml:id="goo19B-FPG_04194-1898.p.1136">
  <s>
    ...
    <w lemma="biti" ana="#Va">je</w><c> </c>
    <w lemma="pripravljati" ana="#Vmp">pripravljal</w><c> </c>
    <w lemma="svoj" ana="#P">svoje</w><c> </c>
    <choice>
      <orig><w>spremljevavce</w></orig>
      <reg><w lemma="spremljevalec" ana="#Ncm">spremljevalce</w></reg>
    </choice><c> </c>
    <w lemma="na" ana="#S">na</w><c> </c>
    <w lemma="žalosten" ana="#Agp">žalostno</w><c> </c>
    <choice>
      <orig><w>iznenadenje</w></orig>
      <reg><w lemma="iznenadenje" ana="#Ncn">iznenadenje</w>
         <desc><gloss>presenečenje</gloss><bibl>SSKJ</bibl></desc>
      </reg>
   </choice>
   <pc>,</pc>
   ...
```

Figure 5 Example from the goo300k hand-annotated corpus.

and, as explained in the following sections, the words assigned linguistic annotation, which was then hand-corrected together with any remaining errors in the text.

The corpus is distributed as a single TEI document (but with each page as an individual file) as well as in derived formats, in particular the so called vertical file for input to CQP/CWB (Christ, 1994) compatible concordancers. The corpus is also mounted on our installation of the noSketchEngine (Rychlý, 2007), a powerful open-source concordancer. The concordancer makes available for searching or display all the word-level annotations, as well as giving metadata for each page including links to the appropriate page in the digital library. The concordancer supports REST-like direct queries via URL, which enables linking other Web-accessible IMP resources, in particular the on-line dictionary entries (cf. Section 6) to corpus query results.

5.1 Word-level modernisation

The basic idea of the linguistic annotation is that (mostly) individual word tokens are modernised where they differ from the contemporary norm, by giving them the inflected word form in contemporary orthography. This modernised form is then assigned its lemma and part-of-speech (PoS) tag. If the orthographically modernised word is archaic in the sense that it now has a different PoS tag (e.g., it changed its gender), exhibits a significant semantic shift or simply does not exist anymore, then the closest modern equivalent(s) or a short gloss is also provided.

As illustrated in Figure 5, each word is annotated, as is whitespace and punctuation. If a word is the same as today then it is simply given its lemma and PoS tag as attributes; where it needs to be modernised, this is encoded by the `choice` tag, which contains the original and regularised forms; the annotations are given to the regularised form, as it is this word form that is lemmatised and tagged, and, in cases of

archaic words, also assigned the gloss and (very basic) bibliographic information giving the source of the gloss. It should be noted that this encoding also supports cases where the historical tokens do not correspond 1:1 with contemporary word forms, as long as the sequences are contiguous ones: the original or regularised form simply contains more than one word. While this approach would not be appropriate for languages such as German, it is sufficient for Slovene, where e.g., there are no separable verbs.

5.2 PoS tagging

While it would be useful to have a fully grammatically annotated historical corpus of Slovene, this is a very labour intensive task. For corpora of contemporary Slovene the MULTEXT-East tagset (Erjavec, 2012a) of the morphosyntactic descriptions (MSDs) is typically used. Slovene distinguishes almost 2,000 different MSDs, making manual annotation slow and requiring highly trained annotators. While the priority in IMP was on modernisation and lemmatisation, it is useful to have at least the basic grammatical information reliably assigned to each word. We therefore developed a tagset which is essentially the Slovene MULTEXT-East one but without inflectional and some lexical features and has only 33 tags, which made the annotation much easier and faster.

The specifications for the IMP tagset are encoded in the same way as they are in MULTEXT-East, i.e., as a TEI P5 document with the attribute-value definitions for each category encoded as tables and also localised to Slovene, so the tags, attributes and their values can be expressed either in English or in Slovene. For example, the tag `Asp` can be transformed into `Adjective, Type = possessive, Degree = positive` or to the Slovene tag `Psn`, which then expands to `pridevnik, vrsta = svojilni, stopnja = nedoločeno`. In addition to the XML the specifications are also available as Web-mounted HTML and as a collection of tabular files.

5.3 The process of manual annotation

Annotation of goo300k was performed by a team of students led by a lexicographer and, together with corpus annotation for lexicon extension (described in Section 6), took almost two years. For annotation we used the CoBaLT editor (Kenter et al., 2012), developed in the scope of the IMPACT project, which supports export to TEI and also allows correcting errors in the base text, a necessary functionality, as linguistic annotation will invariably discover left-over OCR errors. However, corrections can only be performed on the level of individual word tokens, which proved problematic, as there are also frequent errors in punctuation, cases where one word was erroneously split into two or vice versa, all impossible to correct in CoBaLT. We compensated for these difficulties by writing a rather complex series of XSLT scripts that convert the CoBaLT TEI export to the official goo300k corpus.

The corpus was annotated in several annotation campaigns, with the first rounds serving to debug the annotation software, write the annotator's manual, cookbook

and FAQ and to train the students. Specifying the guidelines was a difficult process, as is always the case when trying to capture real language in a system of categories and rules. For example, it is difficult to specify exactly when a word should receive a gloss, i.e., be considered archaic, for example in cases when it only exhibits a (slight) semantic shift against current usage or when the word is still used sometimes, but has a (slight) archaic flavour. In short, the annotation of the corpus had to deal with all common issues faced by lexicographers in their work, with the situation complicated by having many annotators, which were mostly undergraduate students rather than professional lexicographers.

Once the corpus had been annotated, there followed another series of corrections, where error-prone cases were re-examined and corrected where necessary. For example, all lemmas that had more than one interpretation (in terms of their PoS and possibly gloss) were re-examined and, if appropriate, unified. The final goo300k corpus is as error-free as possible, to serve as a gold-standard dataset for research on historical Slovene.

6 The IMP lexicon

In the IMPACT project the main focus of our work was to produce a lexicon of historical Slovene which would be useful in an information retrieval context, i.e., would give modernised forms and especially lemmas of old words. The construction of the lexicon was corpus based, i.e., all the word forms in the lexicon had to be attested in historical sources.[10] The method used was to annotate corpora with all the relevant lexical information, namely the modernised form of the historical word form, the lemma, PoS tag, and optional gloss, i.e., the annotations already discussed in the previous section and present in goo300k. With the corpus containing all the needed information, the compilation of the lexicon is then simply a matter of automatically extracting it from the corpus.

However, goo300k is a rather small corpus and yields a small vocabulary. To extend the lexicon, a ten times larger corpus, called foo3M, was extracted from the IMP DL. It contains 4,500 sampled pages (with pages already in goo300k excluded[11]) from 321 texts, giving a corpus of just under 3 million word tokens. The corpus was then, just as goo300k, automatically annotated and imported into CoBaLT, but with all the word forms from goo300k made invisible to the annotators, as these were already present in the lexicon. The new word forms (ignoring hapax words, which are often typos) were then verified. Unlike goo300k, not all the occurrences of these words in the corpus were examined and corrected, as the goal here was to extend the lexicon rather than fully annotate a corpus. The IMP lexicon is thus a conversion of the annotated and manually validated word tokens from the union of goo300k and foo3M.

[10] In IMPACT such lexicons and the corpora they were based on were developed also for other languages, e.g., for Polish (Bień, 2014) and Spanish (Sánchez-Martínez et al., 2013), where a similar approach to to ours was taken.

[11] The aim was to keep goo300k and foo3M disjoint, so the former can be used as the training and the latter as a realistic test set in lexicon related experiments.

While we tried to make the lexicon as error-free as possible, it most likely still contains some mistakes. Namely, our corpus-based methodology means that if possibly hundreds of manually validated and consistent annotations of a particular word form contain a single error, the lexicon will contain two entries: one correct, and one wrong. At the same time, the lexicon has become quite large, so cross-checking it fully would be a major investment.

6.1 Lexicon structure

The lexicon is encoded using the dictionary module of TEI P5, which has many dedicated elements for dictionary entries. The lexicon is also available on-line, as a set of static HTML files produced by a bespoke XSLT from the TEI source. To illustrate its structure, rather than giving the verbose XML encoding, we show in Figure 6 the beginning of a dictionary entry in HTML. Each entry contains the headword (lemma), in our case *izmed* (*among*) with its grammatical properties, here *predlog* (*preposition*), and possibly a gloss; this information also defines an entry, i.e., an entry is uniquely identified with the triplet consisting of the lemma, PoS and (optional) gloss.

The entry is hyperlinked to the concordancer, either showing all the hits on this entry in goo300k or in the complete IMP corpus (cf. Section 7). It is also hyperlinked to two on-line dictionaries, the *Slovar slovenskega knjižnjega jezika* (*Dictionary of Slovene literary language*), commonly known as *SSKJ*, and the *Pleteršnikov slovar* (*Pleteršnik dictionary*), a Slovene to German dictionary from the end of the 19^{th} century. It should be noted that the link to each corpus or dictionary is present only in cases where the resource returns results on the search query.

Next comes a list of all the contemporary word forms used in the corpus. As the preposition *izmed* is a non-inflecting word, it has only one form, identical to the lemma. Each contemporary form then has all its historical varieties listed; the example shows the first three, *is*, *ismed* and *is med*, the last one being an example of a merged form ("združena oblika"). Finally, each historical form has citations[12] from the corpus, together with their bibliographic information, where the title of the work is hyperlinked to the page containing the citation in the IMP on-line digital library, introduced in Section 4.

6.2 The lexicon in numbers

Table 3 shows some quantitative measures of the lexicon. To give a better idea of the distribution of the types of entries in the lexicon, the table has four columns. The XL column gives the numbers for the complete lexicon. However, as a portion of the lexicon was extracted from the fully manually annotated goo300k, it also contains less interesting entries, such as numerals, typos, foreign words and proper nouns, and the L column gives the sizes for the lexicon without these entries. Many words,

[12] Currently, the TEI lexicon contains 9 random examples, and HTML 4. Including all examples would be also possible, but for very frequent words this is a few thousand examples.)

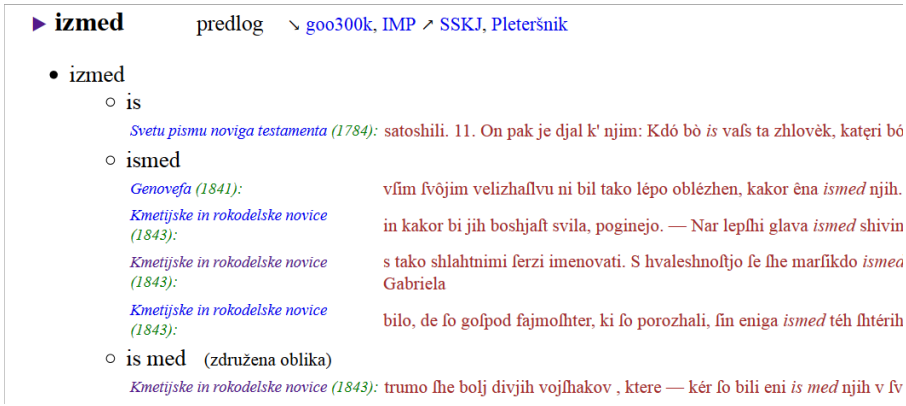

Figure 6 The IMP lexicon in HTML.

	XL		L		M		S	
Entries	28,096		21,653		12,255		4,156	
Lemmas	27,139	1.04	20,917	1.04	11,989	1.02	4,048	1.03
Modern word forms	56,650	2.09	49,212	2.35	24,138	2.01	6,140	1.52
Historical word forms	77,433	1.37	69,326	1.41	34,357	1.42	6,910	1.13
Five-tuples	81,740	2.91	73,263	3.38	35,941	2.93	7,021	1.69

Table 3 Quantitative measures of the IMP lexicon.

esp. in more recent texts, are identical to the contemporary ones; to see how much material that has undergone change is contained in the lexicon, the M column gives only those words (and their lexical entries) that are different from the contemporary ones. Finally, the S column gives only those entries that are archaic, i.e., have a gloss associated with the entry.

The first row gives the number of lexical entries in each sublexicon. As can be seen, about a fourth of the complete XL lexicon contains words, that are not interesting from a linguistic point of view, about half of the L lexicon contains words that have not changed their spelling, while about a third of the M lexicon consists of words that are archaic. The second row gives the number of lemmas, i.e., the headwords in the entries, regardless of their PoS or gloss. The numbers next to the raw counts give the proportion of entries to lemmas, i.e., in how many entries, on the average, the same headword appears. This number gives the ambiguity of the lemmas, which is in the range of 2–4 %. The third row gives the number of contemporary word forms; these, of course, do not cover the complete inflectional paradigms of words, only those that have attested historical variants in the corpora. The second number here gives the proportion of the modern forms to the lemmas, showing how many word forms, on the average, a lemma has; expectedly, the lowest numbers are with M and S lexicons, the first due to the fact that the lexicon contains only changed forms, even though the corpora contain other inflectional forms which have not changed, and the second because the majority of the archaic words come from the oldest parts of the corpus, which are quite small and thus contain only a few occurrences of each lemma. The fourth row gives the number of historical word forms, i.e., words that appear

(modulo capitalisation) in the corpora. The difference between XL and L columns shows that about 10% are linguistically uninteresting word forms, while comparing L and M shows that about half are still with us today, and L and S that about 10 % of linguistically interesting word forms are archaic. The numbers that follow the counts give the proportion of historical to contemporary words, i.e., how many historical forms, on the average, a modern word has in the corpora. In the L lexicon, this is just over 1.4, showing the variability of spelling in the historical sources. Finally, the last row gives the numbers of the complete five-tuples, where each contains all the lexical information on a historical word form, i.e., its contemporary word form, PoS, lemma and gloss; the numbers following the counts give the proportion of the number of entries to the number of six-tuples, i.e., how many different historical words, on the average, are contained in one entry, which is around three for all of the lexica, except the archaic one, where it is, of course, much less.

7 The digital library as corpus

We developed a tool, called ToTrTaLe (Erjavec, 2011), a pipeline which performs tokenisation, transcription, PoS tagging and lemmatisation over input TEI P5 documents and outputs TEI P5 with added in-line linguistic annotation. For historical texts the "transcription" step modernises word tokens by first using a fixed lexicon giving the mappings of historical words to contemporary ones, then a contemporary lexicon — in our case the Slovene reference lexicon Sloleks[13] (Arhar, 2009) — for words that are already in their contemporary form, and, finally, for out-of-vocabulary (OOV) words, a set of transcription patterns which were developed in the course of the IMPACT project. These patterns are applied non-deterministically by the Vaam finite-state library (Reffle, 2011) and their output filtered against the (Sloleks) lexicon of contemporary words.

Figure 7 gives as an example the start of the text presented in Figure 3 but this time annotated with ToTrTaLe. The linguistic annotation is the same as in the manually annotated corpus, except that here the PoS tagging uses the full Slovene MULTEXT-East tagset rather than the simplified IMP one; the reason for this is that the tagger, once the words have been modernised, uses the tagging model developed for contemporary Slovene with its full tagset. The other point to note is that the text in the `figDesc` element is not analysed; the reason for this is that the figure description is not part of the base text but rather meta-data, and thus should not be subject to analysis. ToTrTaLe is TEI-aware and therefore selects the text that should be analysed.

The initial version of ToTrTaLe was used to annotate the base goo300k and foo3M corpora, which were then manually corrected, while the current version uses the complete IMP lexicon as the background lexicon. We performed an evaluation of the tool (Erjavec, 2014) where we used goo300k to construct a background historical lexicon with the test set being the lexicon extracted from the foo3M corpus. The achieved accuracy over OOV lexicon entries (so, per word form type) varies with the age of the text, from 49.8% for the lexicon from the second half of the 18^{th} century to 86.9% for the second half of the 19^{th} century.

[13] Sloleks is available for download under CC BY-NC from `http://eng.slovenscina.eu/`.

```
<body>
  <pb n="[0]" facs="#FPG00008-000" xml:id="pb.001"/>
     <figure xml:id="figure.1">
        <figDesc>Svetnik s kelihom v levi roki, na kelihu kača,
                 desno roko ima rahlo privzdignjeno</figDesc>
      </figure>
      <p xml:id="p.1">
        <s>
          <choice>
            <orig><w>BRATERNA</w></orig>
            <reg><w lemma="bratrina" ana="Ncfsn">bratrina</w></reg>
          </choice>
          <c> </c>
          <choice>
            <orig><w>SVETIGA</w></orig>
              <reg><w lemma="svet" ana="Agpmsg">svetega</w></reg>
          </choice>
          <c> </c>
          <w lemma="janša" ana="Npmsn">JANŠA</w>
         <pc>.</pc>
       </s>
     </p>
```

Figure 7 Example of a ToTrTaLe analysed text.

The complete IMP DL was recently annotated with ToTrTaLe and mounted (along with goo300k, as already mentioned in Section 5) on our installation of noSketchEngine, so a significant part — almost 15 million words — of Slovene written cultural heritage is now available on-line for linguistic research.

8 Conclusions

The paper presented the IMP language resources of historical Slovene,[14] which comprise a TEI P5 encoded data-set consisting of facsimiles with aligned proof-read and structurally annotated texts, a gold-standard manually annotated corpus, and a lexicon. The resources can serve for research and development in language technologies, e.g., for better OCR, information retrieval and ease of reading of historical Slovene. While not the goal of the supporting projects, we also tried to make the resources available in a digital humanities context, including teaching, in the form of a digital library, also accessible through a Web-based concordancer for diachronic corpus linguistics, and a hyperlinked dictionary.

The IMP resources have come about by a lucky confluence of several projects, all dedicated to historical language, but each one with its own slant and also focusing on a different type of printed materials. The result is a much richer, varied and larger set of resources than is usually the case in similar projects. On the other hand, the production was much more demanding, as it was necessary to balance the competing requirements and time frames of the projects, as well as trying to extend their goals

[14] The home page of IMP is `http://nl.ijs.si/imp/`.

into digital humanities. Still, the produced resources are quite extensive for a language of only 2 million speakers. Compared to the Oxford Text Archive[15], which is the oldest archive of out-of-copyright texts (and, in many ways, the home of the TEI) and currently offers 2,723 TEI encoded texts, the IMP digital library has 658, i.e., almost a quarter of the number, and this for a language that has a miniscule fraction of text production and speakers compared to English.

To evaluate the usefulness of the IMP resources and gain insight into future requirements we recently performed a usage study of the resources (Erjavec and Fišer, 2014). The HTTP server logs from 2013 and first half of 2014 for the digital library, on-line dictionary and concordancer over the IMP corpora showed 2,000–3,400 accesses per month (excluding robots), or about 100 per day. We also performed an on-line survey completed by 52 respondents, two thirds of which are elementary and secondary school teachers, followed by under- and postgraduate students. Two thirds use the resources about once a month, with the rest equally divided into more and less frequent users. Almost all respondents (over 90%) agree with the statements that the resources are useful for linguistic research, for teaching Slovene language, and, in general, important for the Slovene society. On the critical side, almost half (about 40%) state that the resources should be enlarged and that searching is too complex. Somewhat fewer thought that the visualisation should be improved, while only 17% are bothered by errors. Almost all (95%) think that the resources are worthy of recommendation and 35% have already done so. We were also happy with the response to where the main focus of future improvements to the resources should be, as the answers were almost uniformly spread between giving priority to the needs of linguists (14%), pupils and students (both at 13%), elementary and secondary school teachers (both 12%) and lexicographers (9%), showing that resources (could) serve a wide range of users.

In our future work we plan to extend the scope of the IMP texts by incorporating the latest additions to the Wikivir project and other available transcriptions, esp. of manuscripts, as these are also interesting for future handwriting recognition research. We plan to incorporate into the corpus the eZISS[16] text-critical editions of Slovene literature, where carefully transcribed older manuscript materials are available. We could also make the digital library and dictionary more user friendly, by e.g., making the transcriptions available also in other formats, such as PDF and ePub, and by mounting them under a fully-fledged DL platform. However, this brings with it the question whether our institute is the right place to host a proper and maintained digital library; a better option would be to incorporate the IMP DL into dLib.si but this would involve significant software upgrade and investment by the National and University Library of Slovenia, i.e., the migration is dependent on dedicated project funding.

While we have already implemented the ToTrTaLe annotation tool, this should be taken only as a preliminary step in the development of automatic annotation methods for historical Slovene. With the resources in place, it is now possible to experiment with inductive methods for word form modernisation. In follow-up work (Scherrer and Erjavec, 2013) we have applied character-based statistical machine translation

[15] `http://ota.ahds.ac.uk/`

[16] The eZISS digital library home page is `http://nl.ijs.si/e-zrc/`.

(CSMT) for the task of transcription, and the initial results are significantly better than when using hand-constructed rules. It should be noted that the experimental data (training and testing dataset) used with the CSMT modernisation have also been made freely available and published along with the paper, so that others could test their methods of word modernisation on our data.

In further experiments we would like to pursue two research directions. On the one hand, modernising old Slovene language is not much different from normalising any other non-standard language. We have already applied the CSMT approach to normalising Slovene tweets (Ljubešić et al., 2014) with encouraging results: if the lemmatisation accuracy over original tweets with the model trained on standard language gives only 75.0% accuracy, it rises to 83.6% if the words are first normalised with our CSMT model. In addition to Slovene user generated content we plan to apply similar methods also to transcriptions of spoken data, given that there now exists GOS, a large and CC available corpus of spoken Slovene (Verdonik et al., 2013).

Finally, we plan to develop a Web application which would modernise old texts or age contemporary texts and make it available on the Web and for mobile phones, in order to familiarise pupils and students with historical Slovene.

Acknowledgements The author thanks the two anonymous reviewers for useful comments and suggestions. For collaborating in the compilation of the IMP language resources thanks are due to Kozma Ahačič, Simon Atelšek, Tina Benčina, Katja Cingerle, Metod Čepar, Darja Fišer, Miran Hladnik, Alenka Jelovšek, Urška Kamenšek, Alenka Kavčič Čolić, Domen Kermc, Maša Kodrič, Simon Krek, Nina Mikulin, Matija Ogrin, Daša Pokorn, Erich Prunč, Zala Šmid, Ines Vodopivec and Maja Žorga Dulmin. The work presented in this paper was supported by the Austrian Academy project "Deutsch-slowenische / kroatische Übersetzung 1848-1918", the EU IMPACT project "Improving Access to Text", the Google Digital Humanities Research Award "Language models for historical Slovenian", and the Research Programme P2-0103 "Knowledge Technologies" funded by the Slovenian Research Agency.

References

Arhar Š. (2009) Učni korpus SSJ in leksikon besednih oblik za slovenščino (The SSJ training corpus and word form lexicon for Slovene). Jezik in slovstvo 54(3–4):43–56

Bień J.S. (2014) The IMPACT project Polish Ground-Truth texts as a DjVu corpus. Cognitive Studies | Études Cognitives (14):75–84, URL `http://bc.klf.uw.edu.pl/381/`

Christ O. (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, Budapest, Hungary, pp. 23–32.

Clausner C., Pletschacher S., Antonacopoulos A. (2011) Aletheia – an advanced document layout and text ground-truthing system for production environments. IEEE Xplore Digital Library pp. 48–52.

Dudczak A., Kmieciak M., Werla M. (2012) Creation of Textual Versions of Historical Documents from Polish Digital Libraries. In: Lecture Notes in Computer Science, Springer, vol. 7489, pp. 89–94.

Erjavec T. (2007) An Architecture for Editing Complex Digital Documents. In: Proceedings of INFuture'07 "Digital Information and Heritage", University of Zagreb, pp. 105–114.

Erjavec T. (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, Association for Computational Linguistics, Portland, OR, USA, pp 33–38, URL http://www.aclweb.org/anthology/W11-1505

Erjavec T. (2012a) MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. Language Resources and Evaluation 46(1):131–142.

Erjavec T. (2012b) The goo300k corpus of historical Slovene. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey.

Erjavec T. (2014) Posodabljanje starejše slovenščine (Modernising historical Slovene). Uporabna informatika 21(4):186–195.

Erjavec T., Fišer D. (2014) Recepcija virov starejše slovenščine IMP (The reception of the IMP historical language resources). In: 33. simpozij Obdobja, Znanstvena založba Filozofske fakultete, Ljubljana.

Erjavec T., Vodopivec I., Kodrič M. (2011) Izdelava korpusa starejših slovenskih besedil v okviru projekta IMPACT (The compilation of a corpus of historical Slovene texts in the scope of the IMPACT project). In: 30. simpozij Obdobja, Znanstvena založba Filozofske fakultete, Ljubljana, pp. 121–127.

Hladnik M. (2009) Infrastruktura slovenistične literarne vede (The infrastructure of Slovene literary studies). In: 28. simpozij Obdobja, Znanstvena založba Filozofske fakultete, Ljubljana, pp. 161–169, URL http://www.centerslo.net/files/file/simpozij/simp28/Hladnik.pdf

Kenter T., Erjavec T., Žorga M., Fišer D. (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, ACL, Avignon, France.

Krauwer S. (2003) The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap. In: Proceedings of the International Workshop Speech and Computer (SPECOM 2003), Moscow State Linguistic University, pp. 8–15, URL http://www.elsnet.org/dox/krauwer-specom2003.pdf

Kroch A., Santorini B., Diertani A. (2004) Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/

Kučera K. (1999) The General Principles of the Diachronic Part of the Czech National Corpus. In: Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 1692, Springer Berlin / Heidelberg, pp. 841–842.

Ljubešić N., Erjavec T., Fišer D. (2014) Standardizing Tweets with Character-level Machine Translation. In: Gelbukh A (ed) 15th International Conference, CICLing 2014, Proceedings, Part II, Springer, Lecture Notes in Computer Science, vol. 8404, pp. 164–175.

Piotrowski M. (2012) Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Pletschacher S., Antonacopoulos A. (2010) The PAGE (Page Analysis and Ground-truth Elements) Format Framework. In: Proceedings of the 20th International Conference on Pattern Recognition (ICPR), Istambul.

Prunč E. (2007) Deutsch-slowenische/kroatische Übersetzung 1848-1918. Ein Werkstättenbericht. Wiener Slavistisches Jahrbuch (53):63–176.

Rayson P., Archer D., Baron A., Culpeper J., Smith N. (2007) Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English Corpora. In: Corpus Linguistics Conference (CL2007), University of Birmingham, Birmingham, UK, URL http://ucrel.lancs.ac.uk/publications/CL2007/paper/192_Paper.pdf

Reffle U. (2011) Efficiently generating correction suggestions for garbled tokens of historical language. Natural Language Engineering 17:265–282.

Rychlý P. (2007) Manatee/bonito – a modular corpus manager. In: Proceedings of 1st Workshop on Recent Advances in Slavonic Natural Language Processing, Masaryk University, Brno, pp. 65–70.

Sánchez-Marco C., Boleda G., Fontana J.M., Domingo J. (2010) Annotation and Representation of a Diachronic Corpus of Spanish. In: Proceedings of the Seventh Conference on Language Resources and Evaluation (LREC'10), ELRA, Valletta, Malta.

Sánchez-Martínez F., Martínez-Sempere I., Ivars-Ribes X., Carrasco R.bbC. (2013) An open diachronic corpus of historical spanish. Language Resources and Evaluation 47(4):1327–1342.

Scheible S., Whitt R.J., Durrell M., Bennett P. (2011) A Gold Standard Corpus of Early Modern German. In: Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics, Portland, Oregon, USA, pp. 124–128, URL http://www.aclweb.org/anthology/W11-0415

Scherrer Y., Erjavec T. (2013) Modernizing historical Slovene words with character-based SMT. In: BSNLP 2013 - 4th Biennial Workshop on Balto-Slavic Natural Language Processing, Sofia.

TEI Consortium (ed.) (2012) Guidelines for Electronic Text Encoding and Interchange. TEI Consortium, URL http://www.tei-c.org/P5/

Verdonik D., Kosem I., Vitez A.Z., Krek S., Stabej M. (2013) Compilation, Transcription and Usage of a Reference Speech Corpus: The Case of the Slovene Corpus GOS. Language Resources and Evaluation 47(4):1031–1048.

Wallenberg J., Ingason A.K., Sigurthsson E.F., Rögnvaldsson E. (2011) Icelandic Parsed Historical Corpus (IcePaHC), Version 0.9. URL http://www.linguist.is/icelandic_treebank