A lexicon for processing archaic language: the case of XIXth century Slovene

Tomaž Erjavec[1], Christoph Ringlstetter[2], Maja Žorga[3], Annette Gotscharek[2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
[2] Centre for Language and Information Processing, University of Munich
Schellingstrasse 10, 80799 Munich
kristof@cis.uni-muenchen.de, annette@cis.uni-muenchen.de
[3]maja.zorga@gmail.com

Abstract

The paper presents a lexicon to support computational processing of historical Slovene texts. Historical Slovene texts are being increasingly digitised and made available on the internet but are still underutilised as no language technology support is offered for their processing. Appropriate tools and resources would enable full-text searching with modern-day lemmas, modernisation of archaic language to make it more accessible to today's readers, and automatic OCR correction. We discuss the lexicon needed to support tokenisation, modernisation, lemmatisation and part-of-speech tagging of historical texts. The process of lexicon acquisition relies on a proof-read corpus, a large lexicon of contemporary Slovene, and tools to map historical forms to their contemporary equivalents via a set of rewrite rules, and to provide an editing environment for lexicon construction. The lexicon, currently work in progress, will be made publicly available; it should help not only in making digital libraries more accessible but also provide a quantitative basis for linguistic explorations of historical Slovene texts and a prototype electronic dictionary of archaic Slovene.

1. Introduction

A large number of Slovene books and periodicals from the XIXth century and earlier are being made available on the internet, e.g. via the dLib.si digital library (Krstulović and Šetinc, 2005), the Slovene literary classics project at WikiSource and Google Books.[1] Human language technology support could bring increased functionality to such digital libraries, esp. for full-text search and information retrieval. The most obvious task is automatic lemmatisation of text, which abstracts away from the morphological variation encountered in heavily inflecting languages, such as Slovene. The user can thus query for e.g. *mati* (*mother*) and receive portions of text containing this word in any of its inflected forms (*matere, materi, materjo*, etc.). Support for lemmatisation, as well as morphosyntactic tagging is well-advanced for modern-day Slovene (Erjavec & Krek, 2008). However, the situation is very different for historical Slovene, where no such research has yet been carried out for the language.

Historical Slovene[2] 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 texts older than 1850 which use the Bohoričica alphabet.[3]

We are currently developing a tool-chain for processing archaic Slovene texts which should alleviate some of these problems. The tool, called ToTrTaLe, is an extension of the ToTaLe tool (Erjavec et al., 2005), which performs tokenisation, tagging and lemmatisation, but extended with a transcription module: after tokenisation, the word-forms are first modernised as regards spelling, and only then passed on to the tagging and lemmatisation modules. This approach follows Rayson et al. (2007) in being able to use the well-developed tagging (and lemmatisation) models for contemporary language rather than having to first develop such models for historical language – a very lengthy and expensive process.[4] The approach has the further benefit of offering the contemporary words paired with archaic ones.

This paper focuses on the transcription aspect of this process which crucially depends on a lexicon or, rather, a series of lexica for the language. In previous work (Erjavec et al., 2010) we concentrated on the first steps

[1] Hladnik (2009) gives a good overview of digitisation efforts and availability of Slovene texts on the internet.

[2] In this paper we concentrate on the Slovene from the XIXth century; the problems are, of course, worse going further back in time, but even here, due to the late development of the written Slovene word and its spelling standardisation, there are substantial differences to contemporary Slovene.

[3] The Bohoričica alphabet had different conventions in writing various Slovene sounds, e.g. »*shaloſt*« is the modern-day »*žalost*«, which makes it confusing for today's readers. Of course, there are also substantial vocabulary as well as syntactic differences, to contemporary Slovene.

[4] For example, annotating for lemma and morphosyntactic description 300,000 words of contemporary Slovene (Erjavec et al., 2010) took about 1,500 hours of annotator time.

(tools and work-flow) involved in manually producing a lexicon of historical Slovene. In this paper we report on the already developed lexica as used in the context of ToTrTaLe.

The rest of the paper is structured as follows: Section 2 details the process of transcription, Section 3 describes the corpora we use in our work, Section 4 the lexica that are used and being produced, Section 5 the silver-standard lexicon, to be made publicly available, Section 6 an experiment studying the current coverage of ToTrTaLe and Section 7 gives some conclusions and directions for further work.

2. Transcription

In this section we explain how modern-day equivalents are found for words in the historical texts, as this represents the main difference to processing modern-day language. The process relies on three resources:

1. A lexicon of modern-day word-forms with associated lemmas and morphosyntactic descriptions.
2. A lexicon of archaic word-forms, with associated modern-day equivalent word-form(s)[5].
3. A set of transcription patterns, giving mappings for changes in alphabets (transliteration) and common spelling changes.

In processing historical texts, the word-forms are first normalised, i.e. de-capitalised and diacritic marks over vowels removed; the latter is most likely Slovene specific, as modern-day Slovene, unlike the language of the 19th century, does not use vowel diacritics.

The following filtering steps are performed on the normalised word-form: if the normalised word-form is an entry of the archaic lexicon, the equivalent modern-day word-form has also been identified; if not, it is checked against the modern-day lexicon. Obviously, if the normalised word-form is found in the modern-day lexicon, its modern-day equivalent has been ipso-facto found as well. This order of searching the dictionaries is important, as the modern lexicon can contain word-forms which have an incorrect meaning in the context of historical texts, so the historical lexicon also serves to block such meanings. For example, the auxiliary verb form *sem* used to be written as *sim* – but in the modern lexicon this is identified as a noun, i.e. the SIM card of a mobile telephone.

If neither lexicon contains the word, the transcription patterns are tried. Many historical spelling variants can be traced back to a set of rewrite rules or "patterns" that locally explain the difference between the contemporary and the historical spelling. For Slovene, e.g., a very prominent pattern is *r→er* as exemplified by the pair *brž→berž*, where the left side represents the modern and the right the historical spelling. Patterns can also be sensitive to the word boundary, as some spelling changes occur only at the start or the end of the word, e.g. *žganjem→žganjam*, where the inflectional ending *-am* has changed into modern-day *-em.* To enable this functionality the appropriate patterns make use of the special symbol, "@", e.g. *em@→am@.*

By corpus inspection we have currently developed a set of about 100 such patterns. These patterns are operationalized by the finite-state tool Vaam (Variant aware approximate matching). Vaam (Reffle, 2011) takes as input a historical word-form, the set of patters, and a modern-day lexicon and efficiently returns the modern-day word-forms that can be computed from the archaic one by applying one or more patterns; the output list is ranked, preferring candidates where a small number of pattern applications is needed for the rewrite operation. Vaam also supports approximate matching based on edit distance, useful for identifying (and correcting) OCR errors; we have, however, not yet made use of this functionality.

It should be noted that the above process of transcription is non-deterministic. While this rarely happens in practice, the historical word-form can have several modern-day equivalents. More importantly, the Vaam module will typically return several possible alternative modernisations. We currently determine the "best" transcription by choosing the most frequent contemporary word between the possible modernisations, but more advanced models are possible, which postpone the decision of the best candidate until the tagging and lemmatisation has been performed.

3. Corpora for lexicon building

To support our work on lexicon acquisition, we use several corpora of Slovene; this section gives the details of the corpora and briefly describes the concordancer used for their inspection.

3.1. Modern language corpora

For lexicon construction, including comparative studies of historical language as opposed to modern language, contemporary corpora are needed. For this purpose we are using several corpora, all based on the FidaPLUS[6] reference corpus of modern Slovene (Arhar and Gorjanc, 2007). FidaPLUS contains 600 million words, where the words have been automatically annotated with morphosyntactic tags and lemmas. The corpora we are using are the following, with the first two having been developed in the JOS[7] project (Erjavec et al., 2010):

- jos100k is a 100,000 word sampled corpus of modern Slovene, with carefully hand-validated word-level morphosyntactic and lemma annotations
- jos1M is ten times larger than jos100k but has only partially hand-validated annotations
- fpj100M is a 100 million sample from FidaPLUS, and has only automatically assigned annotations.

[5] The two lexica have in fact a somewhat more complicated structure, which is further addressed in Section 4.

[6] http://www.fidaplus.net/

[7] http://nl.ijs.si/jos/

These three corpora thus enable studying lexical phenomena choosing either very accurate annotations, but small dataset, or vice-versa. Which option is best depends to a high degree on the frequency of the phenomenon (lexica item) being inspected.

3.2. Historical language corpora

The corpus of historical language we have been mostly using so far was compiled in the scope of the project *Deutsch-slowenische / kroatische Übersetzung* 1848–1918 (Prunč, 2007). The project addressed the linguistic study of Slovene and Croatian books translated from German in the period 1848–1918, where a large portion of the effort went towards building a digital library (compiling a corpus) of the Slovene translations. To this end, the books were first scanned and OCRed, and then, for a portion of the corpus, the transcription was hand-corrected, marked-up with structural information, and, for a few books, lemmatised; this process was supported by a web interface (Erjavec, 2007).

The sub-corpus chosen for building the historical lexicon includes all the AHLib proof-read books written before the year 1900, where the oldest one was published in 1847. There are all together 71 such books, of which the majority (56) are fiction (mostly novels) while 15 are non-fiction (from self-help books for farmers, to text-books on astronomy, chemistry, etc.). All together the corpus contains approximately 2.2 million running words. While certainly small compared to most corpora of contemporary language, it is large and varied enough to have enabled us to start building the historical lexicon.

Recently, we have also collected the older materials available from the WikiSource Slovene literary classics project,[8] led by Prof. Miran Hladnik from the Ljubljana University. In the scope of this on-going project, the raw OCR of books and other materials is being hand-corrected by students. We have downloaded the currently finished transcriptions and turned them into a uniformly encoded corpus. Due to the lack of conventions in structuring Wiki entries, the quality of the automatically acquired meta-data is not very high, however, the corpus makes up for this lack by its size: our current WikiSource corpus contains over 500 publications with over 8 million words. This corpus contains, in general, more recent texts than AHLib, most from the late 19th and early 20th century.

Further historical materials are currently also being hand-corrected, which are meant to extend the scope of the corpus, currently still lacking materials from the 18th century, further into the past.

3.3. The concordancer

All the collected historical corpora are being processed by the (current version) of the ToTrTaLe tool and are then, together with the three corpora of contemporary language, made available via a dedicated Web corpus query interface, with CWB (Christ, 1994) as the backend. The concordancer enables searching and viewing the tokens, their normalised and modernised form, the used transcription pattern, and their computed morpho-syntactic description (i.e. fine-grained PoS tag) and lemma, where the view can be either Keyword in Context (KWIC) or a frequency list. The concordancer has proved to be very helpful in determining the status and preferred annotation of the historical lexical items.

4. Types of lexica

This section gives the various types of lexica used by the program, namely: lexicon of contemporary language; historical word-forms with transcriptions into contemporary language equivalents; historical words without contemporary equivalents; words missing in the contemporary language lexicon; abbreviations; and words which need to be re-tokenised in the modernisation step.

4.1. Contemporary language

The lexicon of contemporary Slovene used was extracted from the FidaPLUS corpus, where each word was automatically annotated with its morphosyntactic description (MSD) and lemma. The MSDs are compact strings that represent the morphosyntactic features of the word form, and can be decomposed into features, e.g. the MSD `Ncms` is equivalent to `Noun, Type = common, Gender = masculine, Number = singular.`

The lexicon was gathered from the corpus by extracting all the triplets consisting of the word-form, lemma and MSD. The word-forms were lowercased. Using regular expressions, entries with anomalous "words" were removed, and only those lexical items with a frequency greater than 4 were retained. With this we arrived at a lexicon, which contains about 600,000 word-forms and 200,000 lemmas.

The lexicon is large enough to cover the majority of contemporary lexis found in historical texts, i.e. it has good recall – however, its precision is relatively low, as it contains many false friends. One example (*sim)* was already mentioned; another case is *serca,* an archaic form for *srca* (*heart*$_{[sg,gen]}$), with the lemma *srce.* This form exists in the modern lexicon, but with the lemma *serec* (*horse of a grey colour*). Such word-forms have to be added to the historical lexicon, with the correct interpretation, in order to block them being retrieved from the modern lexicon.

4.2. Historical to contemporary transcriptions

The second lexicon being developed is that of manually verified historical word-forms. The approach is corpus driven, so far using the AHLib corpus, and relaying on LeXtractor (Gotscharek et al., 2010), a specialised editor for historical lexica.

[8] http://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika

▸ **dervi** ← drvi/Vmp drvmi/Ncn ↑

LEMMA	drviti
PART-OF-SPEECH	Verb, Type=main, Aspect=progressive
DERIVATION	drvi + (r_er,2)

ATTESTATIONS:

Nevidne me v prepir ženejo sile , Viharno proč me bitve slà *dervi* ; Vojaški klic zadeva sluhe moje , Se spenja konj , in bojna	*1848 Divica Orleanska*
planini in dričáje se narase kakor gora vse podere in sabo *dervi* . Na zadnje se pa med strašnim treskom razpraši ; ravno tako	*1853 Divji hunci pred mes...*
koleno se mi v grozi trese , — Kam vertoglavna derznost vas *dervi* ? Poznate zemljo to ? Al veste , kako Se snujejo reči na	*1861 Maria Stuart*

LEMMA	drva
PART-OF-SPEECH	Noun, Type=common, Gender=neuter
DERIVATION	drvmi + (r_er,2)(mi@_i@,4)

ATTESTATIONS:

blizo dva črevlja globoko jamo , ktero je po verstah s suhimi *dervi* ino kremenjem napolnil . Potle je derva užgal , ino lamico	*1849 Robinson mlajši*
ponaša vse Primorsko s preobilnim zidnim kamenjem , ktero je z *dervi* vred že od nekdaj bilo poglavitno blago , prevažano po kupčii	*1861 Kratek popis Cesarst...*
umetnosti in šolo za muziko . Mesto Krakov kupčuje z žitom , *dervi* , soljo , vinom , platnom in svinjami . Skoz Krakov se dovaža	*1861 Kratek popis Cesarst...*

Figure 1. Example of a lexical entry in the historical dictionary

LeXtractor incorporates the Vaam pattern matching functionality and supports both a frequency based selection of entries to be added to the lexicon, as well as directly annotating word tokens in corpora. As mentioned, we give details of the manual lexicon building procedure, as well as how LeXtractor was adapted for Slovene, in Erjavec et al. (2010). Here, we will concentrate on the current structure and content of the lexica.

The lexicon we are developing has a simple structure, where each entry contains the following fields:

1. a word-form that has been witnessed in a proof-read historical text
2. the equivalent word-form from contemporary Slovene, possibly together with the patterns which map the former into the latter
3. the contemporary lemma of the word-form
4. the lexical morphosyntactic properties of the lemma
5. attestations of the word-form in the historical corpus

Figure 1 gives an example of such a lexical entry – the entry is formatted in HTML for the ease of illustration. Note that the historical word-form is ambiguous, i.e. it has two possible modern interpretations.

The intention of this manually collected lexicon is to contain the most frequently occurring archaic words in the texts; we have therefore applied frequency selection of the entries, so that closed class words are extensively covered, as are the most common open class words. We are also including as many as possible of short historical words (up to 5 characters in length) as these most frequently have false friends in the modern lexicon, either directly or via pattern application, as is the case of *sim* and *serca*.

4.3. Words without descendants

The other type of historical lexicon concerns word-forms that are missing a modern-day descendant, i.e. they do not have a corresponding contemporary lemma. For such words, LeXtractor does not currently have the functionality to enter a structured entry, apart from a comment and the attestations. Since we decided it useful to further analyse such entries, we currently enter in the comment space the following information:

1. historical lemma, as it would be written today
2. the closest contemporary Slovene synonym(s)
3. the PoS of the historical lemma
4. the source (dictionary, corpus) on the basis of which the synonyms were chosen
5. potential comments

The reasons that we are adding this information are twofold. First, by providing the "virtual" modern word-form, we are increasing the possibility of a user finding this word, even though unsure about its archaic spelling; similarly, the tagger has a greater chance of assigning the correct MSD to such a word. Secondly, while the lexicon of transcribed words is necessary for computational processing of historical texts, it is, in general, not very interesting for humans, esp. the pattern derived entries. But the words without descendants are exactly those that the modern-day reader will most likely not understand at all. So, as long as they have been identified, it is worthwhile assigning them their near-synonyms and giving the source where further information about them can be found. Such a lexicon could then also represent a prototype "bilingual" historical to modern dictionary, which is still lacking for Slovene.

4.4. Missing contemporary words

In order to improve the functionality of the tool and the filter cascade, the maintenance of the modern lexicon is crucial. Rather than modifying this lexicon directly, we, as discussed, either block inappropriate modern words by including them in the historical lexicon, or add missing words via a special lexicon. Of course, there will always

be words missing from the lexicon, and it is not our intention to add all possible contemporary words that could appear in historical texts, esp. as both the tagger and lemmatiser are able to handle unknown words. However, certain words have a rather unpredictable morphology, which causes either the tagger or lemmatiser to misinterpret them – when such cases are noticed they are added to the lexicon of missing contemporary words.

Rather than adding word-forms individually, we have implemented a Web application that is able to generate the complete inflectional paradigm given the lemma and part-of-speech. Constructing exact paradigms on the basis of this information is, in the general case, not possible, so the intention is for the lexicographer to automatically construct such a paradigm, and then edit by hand the erroneous word-forms.

4.5. Abbreviations

A lexicon very important for correct tokenisation and sentence segmentation is that of abbreviations. The tokenisation module of ToTrTaLe takes a list of abbreviations, i.e. strings ending with a full-stop, which, however do not (necessarily) end a sentence; furthermore, the period should be taken as a part of the abbreviation token. Historical language uses some abbreviations not present anymore in contemporary language – these are included to the lexicon of historical abbreviations, and then added to the tokeniser resource file. The lexicon also includes for each abbreviation its expanded form(s), although these are not currently used by the program.

4.6. Token translations

There is a final type of satellite lexicon that we use in ToTrTaLe, which is interesting from a computational perspective. In historical Slovene certain words or morphemes were written apart or together, where it is now the other way around. The most prevalent and productive example is the prefix that forms the superlative degree of adjectives: what used to be written *nar boljši* is now *najboljši.* As (word) tokens in text processing represent the basic division of characters into linguistic units, which are then further annotated, having a mismatch between archaic and contemporary Slovene at this level of description is difficult to process and encode; from being a string transcription and classification problem, the mapping of old to new language becomes one of machine translation. This is an interesting problem, esp. as it is by no means confined to historical language varieties; the same phenomenon can be found in contemporary Slovene (and other languages) where, in informal or "badly written" language people often write certain words apart, or run separate words together.

Luckily, in historical Slovene, apart from the superlative prefix, and a few other minor cases, only a well-defined set of closed-class words have changed their tokenisation. The tokeniser used by ToTrTaLe uses various classes of special lexica; one of these covers compounds, and the other "clitics", i.e. where a prefix or suffix should be split from the word, such as *-lo* in Italian. We have identified all (or most) of the closed class compounds and splits, and have also taken all the superlative adjectives found in the AHLib corpus into the compounds list. At least for these latter, this is only a stop-gap measure; in the case of Slovene superlatives, a simple regular expression (*nar* .*) would cover almost all situations; as mentioned, the general case is, however, much more complicated.

The tokenisation lexicon thus contains two types of tokens, those that should be kept as one token (about 400), and those that should be split (10); in processing, these tokens are given special flags, which are retained in the output. Vaam patterns are also needed to modernise such cases, e.g. *@naj→@nar_,* where the underscore represents the space character.

entries	77,783
words	63,447
lemmas	18,940
modern entries	73,736
historical	3,181
no descendant	529
blocked modern	230
abbreviations	63
merged	44

Table 1. The size of the silver standard historical lexicon

5. Silver standard lexicon

From the partial and heterogeneous lexica we created a "silver standard" historical lexicon, which, in addition to the hand-gathered lexica also contains automatically collected "safe" modern words attested in the historical corpus. The AHLib corpus was annotated with ToTrTaLe, and the lemmas of all the contemporary words were verified against a lexicon composed of the lexicon derived from the jos100k corpus and the large Slovene monolingual dictionary SSKJ. If the automatically assigned annotations matched those in this lexicon then the entry was included in the silver standard lexicon. This approach yields highly reliable lexical entries.

Table 1 gives the size of the current lexicon, where an entry is taken to be the 4-tuple (normalised word-form, modern word-form, modern-lemma, PoS/MSD). The main part of the lexicon is contributed by modern words, while the manually collected part of historical forms currently has about 4,000 entries.

The silver standard lexicon is encoded against a slightly enhanced schema of the LeXtractor lexicon dump XML. As illustrated in Figure 2, each entry is given a type, and is headed by the (normalised) word-form. The entry can have several analyses, each giving the modernised form, lemma, PoS, possibly modern near synonyms and attestations. Entries for the tokenisation lexicon are recognised by having a white-space in the word-form or modern derivation.

```
<entry type="no_descendant">
  <wordform>alipak</wordform>
  <note>kontekst</note>
  <analyses>
    <analysis>
      <lemma>alipak</lemma>
      <pos>C</pos>
      <derivations>
        <derivation>ali pak</derivation>
      </derivations>
      <synonyms>
        <synonym>ali</synonym>
        <synonym>ali pa</synonym>
      </synonyms>
      <attestations>
        <attestation
           src="korpora/FPG06523.txt"
           position="58207">
          <pre>— ali na suhi zemlji gdè v
                Ameriki ,</pre>
          <word>alipak</word>
          <post>le na kterem ostrovi , še
                 dozdaj ni vedel</post>
        </attestation>
…
```

Figure 2 XML encoding of the lexicon.

6. Lexicon coverage

We performed an experiment in which we evaluate the coverage of the ToTrTaLe given the current lexicon(s) and pattern set. As AHLib served as the development data-set, we took for the experiment the Wiki corpus and, as the modern-day baseline, the Slovene part of the SPOOK parallel corpus of recently translated novels.[9] Both corpora were annotated with ToTaLe, and the Wiki corpus also with ToTrTaLe. We were interested in how the annotations of the two corpora differ when processed with the same model, and how the historical corpus annotations differ when processed without or with the transcription.

Table 2 gives the number and proportions of annotation classes, depending on the corpus and mode of processing. The first row gives the number of word tokens and (normalised) word types. The number of types, i.e. the size of the lexicon needed to completely cover the corpora, is quite high, but it of course also includes all the typos etc. from the source corpora.

The second row shows how many words were found in the modern FidaPLUS lexicon. The percentage is significantly lower with the Wiki corpus, esp. if we compare the number of types; from 83% with modern text down to 54% with the historical one. The third line gives the number of modern words found in the silver-standard dictionary derived lexicon; again, the number of types drops from 83% to 54%. Comparing the "Modern" and "Dictionary" number of the Wiki corpus processed with and without transcription, we note that the numbers obtained with transcriptions are slightly lower; the reason is that some words from the modern lexicon are, when using transcription, blocked by the historical lexicon.

The next line gives the number of unknown words; if Spook has about 16% unknown word types, Wiki without transcriptions has over 45%. With transcription this number drops to 39%, i.e. while we do experience some gain, we are still far from reaching modern-day recognition rates. The decrease of unknown words when using transcription can be mostly attributed to the use of patterns; they help in recognising almost 6% of word types, which is, however, only 0.5% of word tokens; and even here we have to take into account that there is no guarantee that the found modern word is in fact the correct one. The rest of the decrease in unknown words is due to the lexicon of historical words. Out of about 4,000 entries currently in the historical lexicon 2,200 were used; this is under 1% of the lexical types, i.e. much less than covered by the patterns, but, conversely, the number of tokens covered by the historical lexicon (0.76%) is greater than that covered by the patterns (0.49%).

Spook	*Tokens*	*%*	*Types*	*%*
Words	1,825,692	100.00	120,723	100.00
Modern	1,773,019	97.11	100,954	83.62
Dictionary	1,708,764	93.60	85,852	71.11
Unknown	52,673	2.89	19,769	16.38
No lemma	920	0.05	584	0.48
Wiki without transcription				
Words	8,219,093	100.00	249,262	100.00
Modern	7,868,823	95.74	135,490	54.36
Dictionary	7,522,562	91.53	109,549	43.95
Unknown	350,270	4.26	113,772	45.64
No lemma	15,796	0.19	5,623	2.26
Wiki with transcription				
Modern	7,858,325	95.61	135,490	54.36
Dictionary	7,512,988	91.41	109,550	43.95
Historical	62,822	0.76	2,231	0.90
Pattern	39,902	0.49	14,560	5.84
Unknown	258,044	3.14	97,732	39.21
No lemma	9,767	0.12	4,398	1.76

Table 2. Coverage of lexica over modern-day SPOOK corpus and 19th century Wiki corpus with and without transcription.

[9] This parallel corpus is being developed in the scope of the SPOOK project, http://lojze.lugos.si/spook/

The last line in all three tables gives that number of words that could not be lemmatised. These words are interesting, as they point to the morphological changes that occurred over time; in the modern corpus there are only 0.5% of such word types, while the Wiki without transcription has 5 times more, 2.26%; transcription lowers this number to 1.76%. Such words which cannot be lemmatised with the model for modern Slovene are very consistently true archaic words, i.e. good candidates for inclusion into the historical lexicon.

7. Conclusions

The paper presented our methodology of building a lexicon to help process historical language, in particular the Slovene of the XIXth century in the context of the ToTrTaLe tool. The background resources of this work are a historical corpus, a contemporary lexicon of Slovene, spelling variation patterns, and the Vaam and LeXtractor software.

In further work we plan to significantly enlarge the historical lexicon; now that the tools have been set-up and we have elaborated the methodology of the lexico-graphical work, we will engage more people to work on the lexicon, with the target size between 10 and 20 thousand entries. We plan to move from the frequency based word selection to annotating corpus tokens directly – this work also connects to our intention of compiling a gold-standard historical corpus with hand validated annotations. Such a corpus is useful for evaluating the precision/recall of various computational annotation methods and underlying resources, say the transcription rules and, of course, the lexicon. As mentioned, we will also extend the corpus with new materials, esp. newspapers and older books.

Current work has also been exclusively empirically driven, i.e. we addressed only issues that directly arose out of the lexical items found in the corpus. In the future we plan to take into account the linguistic research on historical Slovene that has been done so far , as discussed e.g. in Orožen (1996). Hopefully, our computational approach might also reveal new quantitative and qualitative linguistic insights into the language as used in XIXth century Slovenia.

The concordancer to the corpora is already publicly available at http://nl2.ijs.si/ahlib.html. We will also make the produced corpus and lexicon available under a Creative Commons licence, in the hope that it will facilitate further studies of Slovene historical language.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful suggestions. All errors in the paper of course remain our own. The work presented in this paper was supported by the EU FP7 ICT project IMPACT, “Improving Access to Text”.

References

Oliver Christ, 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. Proceedings of COMPLEX ‘94: 3rd Conference on Computational Lexicography and Text Research. 23-32, Budapest, Hungary.

Špela Arhar and Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. [Corpus FidaPLUS: a new generation of the Slovene reference corpus] *Jezik in slovstvo*, 52(2).

Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen, and Ralf Steinberger. Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. In Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland. 2005, pp. 32-36.

Tomaž Erjavec, Simon Krek, 2008. The JOS morphosyntactically tagged corpus of Slovene. In the Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC’08, Paris. ELRA.

Tomaž Erjavec, Christoph Ringlstetter, Maja Žorga, and Annette Gotscharek, 2010. Towards a Lexicon of XIXth Century Slovene. In Proceedings of the Seventh Language Technologies Conference, October 14th-15th, 2010, Ljubljana, Slovenia. Jožef Stefan Institute.

Annette Gotscharek, Ulrich Reffle, Christoph Ringlstetter, Klaus U. Schulz, and Andreas Neumann. Towards information retrieval on historical document collections: the role of matching procedures and special lexica. International Journal on Document Analysis and Recognition, pp. 1-13, 2010.

Miran Hladnik. 2009. Infrastruktura slovenistične literarne vede [Infrastructure of Slovene Literary Studies]. In *Obdobja 28 – Infrastruktura slovenščine in slovenistike*. pp. 161–69.

Zoran Krstulović and Lenart Šetinc. 2005. Digitalna knjižnica Slovenije – dLib.si. [The digital library of Slovenia – dLib.si] *Informatika kot temelj povezovanja: zbornik posvetovanja*, pp. 683-689.

Martina Orožen. 1996. *Oblikovanje enotnega slovenskega knjižnega jezika v 19. stoletju.* [The formation of a unified Slovene literary language in the XIXth Century.] Ljubljana, Filozofska fakulteta.

Erich Prunč. 2007. Deutsch-slowenische/kroatische Übersetzung 1848-1918. Ein Werkstättenbericht. [German-Slovene/Croatian translation, 1848-1918. Workshop report]. *Wiener Slavistisches Jahrbuch 53/2007*. Austrian Academy of Sciences Press, Vienna. pp. 163-176.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicolas Smith, 2007. Tagging the Bard: Evaluating the accuracy of a modern PoS tagger on Early Modern English corpora. In Proceedings of Corpus Linguistics 2007. Uni. of Birmingham, UK.

Ulrich Reffle, Efficiently generating correction suggestions for garbled tokens of historical language, Journal of Natural Language Engineering, Special Issue on Finite State Methods and Models in Natural Language Processing, 2011.