Towards a Lexicon of XIXth Century Slovene

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

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

Abstract

Historical Slovene texts are being increasingly digitized and made available on the internet in the scope of digital libraries, but so far no language-technology support is offered for processing, searching and reading such materials. Appropriate lexical resources for historical Slovene language could significantly increase such support, by enabling better automatic OCR correction, full-text searching and by modernizing archaic language. This paper describes the first steps in creating a historical lexicon of Slovene, which will map archaic word-forms into modern word-forms and lemmas. The process of lexicon acquisition relies on a proof-read corpus of Slovene books from the XIXth century, a large lexicon of contemporary Slovene language, and LeXtractor, a tool to map historical forms to their contemporary equivalents via a set of rewrite rules, and to provide an editing environment for lexicon construction. The envisioned lexicon should not only help in making digital libraries more accessible but also provide a quantitative basis for linguistic explorations of historical Slovene texts.

Prvi koraki v izdelavi leksikona slovenščine devetnajstega stoletja

Čedalje več slovenskih historičnih besedil je digitaliziranih in dostopnih na spletu v okviru digitalnih knjižnic, vendar zaenkrat še ni na voljo jezikovnotehnološke podpore za obdelavo, iskanje in branje takšnih gradiv. Ustrezni leksikalni viri za historično slovenščino bi lahko z omogočanjem popravkov avtomatsko prepoznanega besedila, iskanja po celotnem besedilu in modernizacijo arhaičnega jezika občutno izboljšali tako podporo. Članek opiše prve korake v razvoju historičnega leksikona slovenščine, ki bo pripisal arhaičnim besednim oblikam sodobne besedne oblike in leme. Proces gradnje slovarja se naslanja na korigirani korpus slovenskih knjig 19. stoletja, obsežen leksikon sodobnega slovenskega jezika in orodje, ki omogoča tako preslikavo historičnih oblik v njihove sodobne ustreznice s pomočjo prepisovalnih pravil kot urejevalno okolje za gradnjo slovarja. Tako zastavljeni leksikon ne bo le omogočil večjo dostopnost digitalnih knjižnic, temveč bo predstavljal tudi kvantitativno osnovo za jezikoslovne raziskave historičnih slovenskih besedil.

1. Introduction

In the context of digital libraries human language technology support can bring increased functionality 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 receives 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 & Džeroski, 2004). However, the situation is very different for historical Slovene, where no such detailed research has yet been carried out for the language.

Historical Slovene language[1] 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.[2]

The above problems would be alleviated by using a large lexicon of historical Slovene language giving the mapping of historical word-forms into their modern-day equivalents with associated lemmas. OCR engines could make use of such a lexicon to guide the recognition process; texts could be lemmatised enabling better search; and the texts could be transcribed using modern day equivalents of the word-forms to facilitate reading.

Developing a lexicon of historical Slovene is a very timely undertaking, as a large number of books and periodicals from the XIXth century are being made available on the internet, e.g. in the context of the dLib.si digital library[3] (Krstulović and Šetinc, 2005) and the Slovene literary classics in WikiSource[4] – Hladnik (2009) gives an overview of digitisation efforts and availability of Slovene texts on the internet.

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

[1] In this paper we concentrate on the Slovene language 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.

[2] 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.

[3]http://www.dlib.si/

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

- a word-form that has been witnessed in a proof-read historical text
- the equivalent word-form from contemporary Slovene
- the contemporary lemma of the word-form
- the lexical morphosyntactic properties of the lemma

Compiling such a lexicon with sufficient coverage is a non-trivial process: a representative corpus of proof-read historical texts must be compiled, a comprehensive modern-day lexicon must be obtained, and the word-stock of the former must be matched against the latter. Problems arise in methodological issues (not all historical word-forms even have a modern day equivalent) as well as technological ones (having a good software environment for lexicon construction).

The rest of this paper is structured as follows: in Section 2 we present the language resources we currently use for lexicon construction, in particular the AHLib historical corpus and FidaPLUS contemporary language lexicon; in Section 3 we introduce the LeXtractor software environment used for lexicon construction; Section 4 discusses the main issues so far discovered in building the lexicon; and Section 5 gives some conclusions and directions for further work.

2. The corpus and lexicon

For the historical lexicon of Slovene to be built using the envisaged methodology, three language resources are needed: a proof-read reference corpus of historical texts, a large lexicon of modern-day Slovene, and the patterns of historical spelling variation. In this section we detail the first two resources, and leave the third for the discussion in the next section.

2.1. The AHLib corpus

The corpus we currently use was compiled in the scope of the project *Deutsch-slowenische / kroatische Übersetzung* 1848–1918 (Prunč, 2007). The project addressed the linguistic study of Slovene 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 these 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 subcorpus 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 enable us to start building the lexicon.

2.2. The FidaPLUS lexicon

The lexicon of contemporary Slovene used was extracted from the FidaPLUS corpus[5] (Arhar and Gorjanc, 2007), a large corpus of contemporary Slovene, where each word was automatically annotated with its morphosyntactic description (MSD) and lemma. The MSDs are compact strings that give the morphosyntactic features of the word form, and can be decomposed into features, e.g. the MSD `Ncmsn` is equivalent to the feature set `Noun, Type = common, Gender = masculine, Number = singular, Case = nominative`.

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, and the word-boundary symbol added to the start and end of the word (c.f. next section). Using regular expressions, entries with anomalous "words" were removed, and only those lexical items with a frequency greater than 4 were retained. The MSDs were also reduced to the lexical features, e.g. from `Ncmsan` to `Ncm`, which simplifies the task of the lexicographer when adding new words to the lexicon. With this we arrived at a lexicon, which is large enough to serve as a reference lexicon of modern word-forms. The lexicon contains about 600,000 word-forms and 200,000 lemmas.

3. LeXtractor and approximate string matching

This section first explains the general ideas and principles guiding our corpus-based construction of lexica for historical language and then describes a web tool for collaborative construction of historical lexica, initially conceptualized for German (Gotscharek et al., 2009), and its adaptation for the Slovenian lexicon project.

3.1 Corpus based lexicon construction

Given a sufficiently large historical corpus, we ignore all words found in a contemporary lexicon of the language processed, as well as special contemporary vocabulary such as names, geographic expressions, etc. The remaining words are analyzed by their frequency of occurrence in the historical corpus. This frequency-based construction ensures that the lexicon soon enables a reasonable recall over the word tokens that represent historical variants of contemporary words.

In order to minimize the cognitive load of the lexicographers, we employ a number of advanced NLP techniques. Our intention is the following ideal division of work: it is the role of the machine to produce meaningful suggestions of what to include into the lexicon; the lexicographers are enabled to concentrate on the linguistic decision according to the corpus material; they can just confirm or reject the suggestions. In the real production process (cf. Section 4), difficult cases where more complex actions and unsupported input of the lexicographers is needed also occur. In what follows we describe the resources that are used to come close to the idealistic goal.

Word lists for contemporary vocabulary. To separate between contemporary words and historical spellings, a collection of word lists of contemporary vocabulary is used. We use special lists for names and geographic expressions, as well as a large list D^{mod} that covers the contemporary standard vocabulary of the processed language.

List of patterns. For Slovene as well as for other languages, many historical spelling variants can be traced back to a set of rewrite rules or "patterns" that locally

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

explain the difference between contemporary and historical spelling. The most prominent pattern for Slovene is *r→er* as exemplified by the pair *brž→berž*. Based on corpus inspection and as a side result of lexicon construction, we currently collected a list P of 57 patterns for Slovene; of these, 26 are for transliteration (e.g., *e→ê* or *š→ſh*), and the rest for "proper" changes in spelling. It should be noted that our patters 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 words in D^{mod} are embedded in a special character (@), e.g. *@žganjem@*, and the appropriate patterns make use of this symbol, e.g. *em@→am@*.

Matching modulo patterns. We employ a tool for matching modulo patterns. The tool uses the word list D^{mod} and the list of patterns P as background resources. Given an input token w' occurring in the historical corpus, all entries w in D^{mod} are computed where w' can be obtained from w by applying one or several patterns. The output list is ranked, preferring candidates w where a small number of pattern applications are needed to rewrite w into w'. With each suggestion w the tool also outputs the set of patterns that are used to rewrite w into w'. The tool is implemented as a finite-state device. The lexicon D^{mod} is represented as a deterministic finite-state automaton. For traversal of the automaton, a special procedure has been implemented that takes pattern variation into account, using the list of patterns P.

Lemmatizing contemporary word-forms. The output of the above process are one or several contemporary word-forms w which correspond to a given historical token w'. It remains to assign the correct lemma(s) and part-of-speech (lexical category) to the word-form(s) w. A lemmatiser for the processed language is used to map a contemporary inflected word-form w to all possible corresponding lemmas. In the case of Slovene, "lemmatizing" of w is implemented by the modern Slovene lexicon. The lexicon holds full forms with the lemma and morphological information attached. This enables us to add linguistic features like part-of-speech and morpho-syntactic information to the entry of w'.

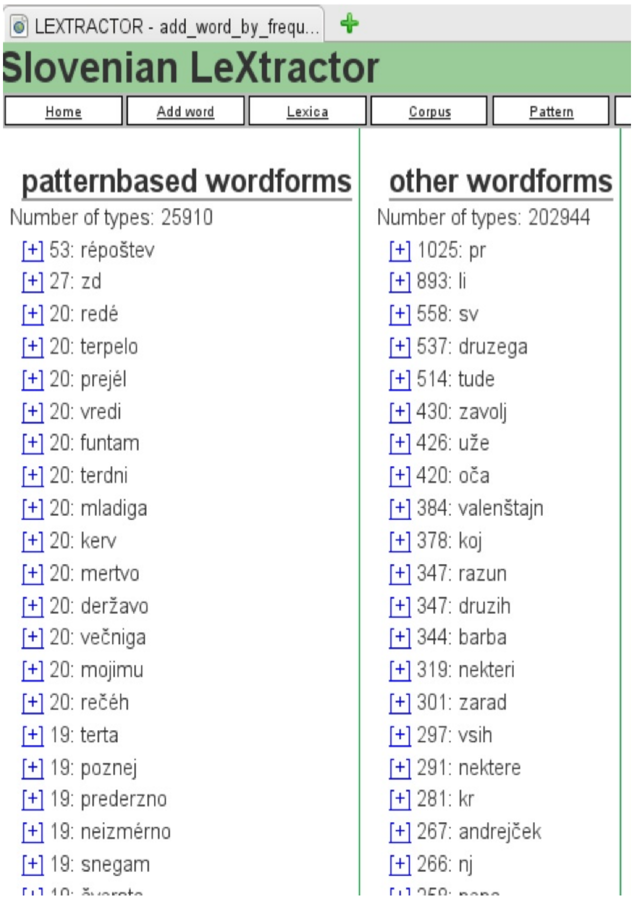
3.2 A web-tool for collaborative construction of lexica for historical language

A web-based tool was designed to implement the workflow of NLP supported collaborative lexicon construction. The main modules of the web-tool are a managing module, which guarantees that no conflicts arise when several lexicographers simultaneously work on the vocabulary of the corpus, an analyzer module, and the graphical user interface; the latter two are described in more detail below.

Given a historical string w' observed in the corpus, the analyzer module first suggests corresponding contemporary word-forms w from the contemporary lexicon D^{mod} based on matching. Each interpretation w comes with the set of patterns that were applied. Second, for a given contemporary word-form w, the analyzer computes all the lemma(s) – including part-of-speech information – which may underlie the word-form w.

The confirmed entries for the historical lexicon are stored in a special database. Standard entries of the database consist of the historical string as found in the corpus, the corresponding contemporary word-form and lemma, the part-of-speech category, pointers to concordances[6] in the historical corpus which serve as attestations for the given interpretation, and the name of the person who created the entry. Note that a historical string can be associated with several entries of the database. The database also contains "non-standard" entries such as named entities, abbreviations, and historical words that do not have a corresponding contemporary lemma.

The graphical user interface visualizes the different workflows to create lexicon entries of the words in the corpus. Figure 1 shows the frequency list mode with the pattern based strings on the left and the non-derivable strings on the right hand side. If the user selects a token w of the left list, she is taken to a new screen that visualizes the possible interpretations of w. By an *interpretation* we mean a pattern based derivation of w from a valid contemporary word-form (cf. Figure 2). The user now confirms or rejects the proposed interpretations. For each confirmed interpretation, the linguistic readings in terms of the corresponding lemma(s) have to be determined.



patternbased wordforms

Number of types: 25910

[+] 53: répoštev
[+] 27: zd
[+] 20: redé
[+] 20: terpelo
[+] 20: prejél
[+] 20: vredi
[+] 20: funtam
[+] 20: terdni
[+] 20: mladiga
[+] 20: kerv
[+] 20: mertvo
[+] 20: deržavo
[+] 20: večniga
[+] 20: mojimu
[+] 20: rečéh
[+] 19: terta
[+] 19: poznej
[+] 19: prederzno
[+] 19: neizmérno
[+] 19: snegam

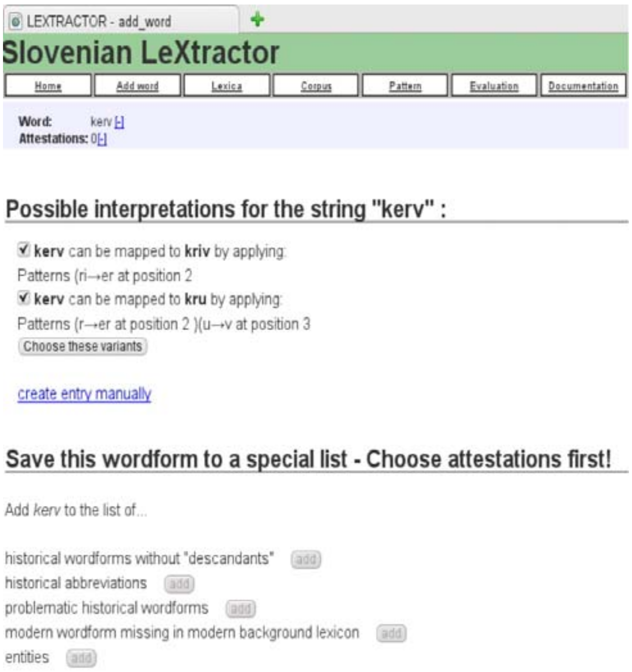
other wordforms

Number of types: 202944

[+] 1025: pr
[+] 893: li
[+] 558: sv
[+] 537: druzega
[+] 514: tude
[+] 430: zavolj
[+] 426: uže
[+] 420: oča
[+] 384: valenštajn
[+] 378: koj
[+] 347: razun
[+] 347: druzih
[+] 344: barba
[+] 319: nekteri
[+] 301: zarad
[+] 297: vsih
[+] 291: nektere
[+] 281: kr
[+] 267: andrejček
[+] 266: nj

Fig. 1 GUI for collaborative lexicon building, corpus mode. Unchecked word-forms derivable by patterns from contemporary words are presented in the left column, ordered by frequency; the non-derivable word-forms are in the right-hand column.

[6] A *concordance* is a text window containing an occurrence of the word form.



Word: kerv [-]
Attestations: 0[-]

Possible interpretations for the string "kerv" :

☑ **kerv** can be mapped to **kriv** by applying:
Patterns (ri→er at position 2

☑ **kerv** can be mapped to **kru** by applying:
Patterns (r→er at position 2)(u→v at position 3

Save this wordform to a special list - Choose attestations first!

Add *kerv* to the list of...

historical wordforms without "descandants" add
historical abbreviations add
problematic historical wordforms add
modern wordform missing in modern background lexicon add
entities add

Fig. 2 Selecting possible interpretations for "*kerv*". The system suggests "*kirv*" and "*kru*". Alternatively, "*kerv*" can be added to special lists visible in the lower part.

In our case, readings based on the contemporary lexicon are suggested by the system. Each reading is confirmed or rejected. Before the lexicographer may confirm a reading she has to select at least one attestation, i.e. a concordance where the reading in question is the correct one. To this end, all concordances are shown graphically (cf. Figure 3).

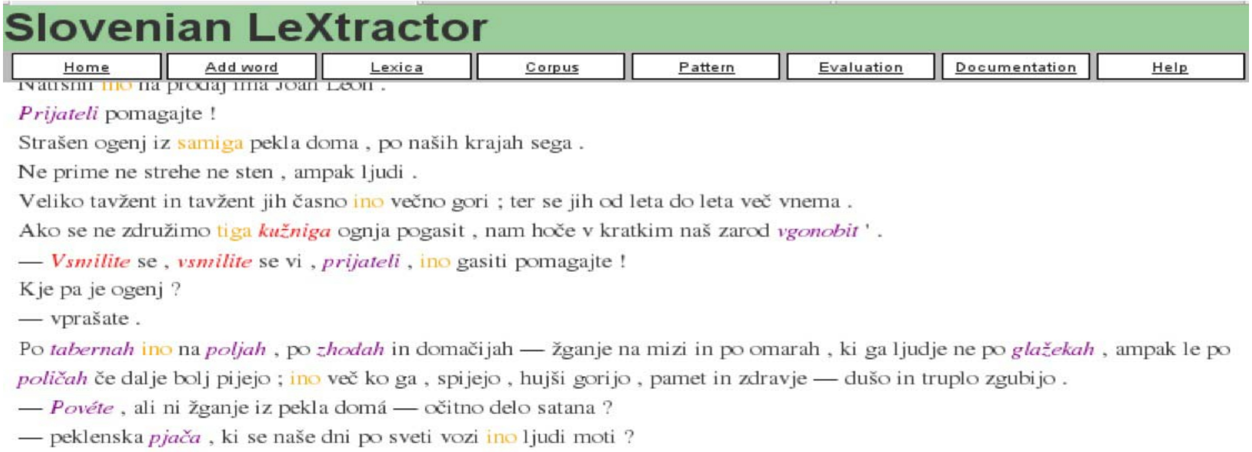
For every confirmed reading, a separate lexicon entry is created that includes the associated attestations. If a processed string has other than pattern based mappings to a contemporary word form or lacks a contemporary explanation, it is included into one of the following *special sublexica*: historic words without a contemporary equivalent; historic abbreviations; historic word-forms which lack a simple transition pattern; named entities; missing words of the contemporary lexicon (cf. Figure 2, lower part).

Entries of the right frequency list are more complicated because they are not rule-based variants in terms of patterns. The system can't suggest the mapping of a historical string *w'* to its contemporary equivalent *w* automatically, so *w* has to be specified manually. If necessary, the lexicographer can also assign these entries to the special lexica mentioned above. If the lexicographer sees a derivation from a contemporary word using a *new* pattern *p* she can suggest to add *p* to the list of patterns. In the current version, there are no automated update mechanisms for the list of patterns and the matching procedure.

Document mode. The lexicographer may also decide to work on a specific text. On the basis of the current lexicon and the matching rules, the text is visualized with all words marked according to their lexical explanation (cf. Figure 3). Additional information is provided through mouse events. We distinguish: contemporary words, checked entries of the lexicon for historical word-forms, entries of the left (right) frequency list shown in the corpus mode, and non-explained strings. If a string is activated in the document mode, the sequential processing is the same as for the corpus mode.

The system is web based and collaborative. Both issues are of great importance for the project. As the involved lexicographers do not work at the same location, flexibility concerning their individual workplaces is needed. Since the professional abilities of the contributors and the complexity of certain lexical entries differ significantly, a workflow was created that leaves more challenging entries in the frequency list to the trained historical linguists, whereas the other lexicographers deal with the simple cases.

From our present perspective, corpus, matching rules, and lexicon should be considered as a joint knowledge base. Given a set of patterns we may use historical word-forms and corresponding contemporary word-forms stored in the lexicon and in addition the corpus for deriving meaningful probabilities or edit weights for the patterns. As a matter of fact, frequency based lexicon construction also helps to find new relevant patterns. In this sense, lexicon and corpus provide empirical evidence for patterns (rules) and help to fine-tune approximate matching. Conversely, we have seen above how refined matching procedures help to speed up lexicon construction. Summing up, this shows that refinement of matching procedures and lexicon construction can be directly interleaved in a kind of bootstrapping procedure.



Natisnil ino na prodaj ima Joan Leon .
Prijateli pomagajte !
Strašen ogenj iz samiga pekla doma , po naših krajah sega .
Ne prime ne strehe ne sten , ampak ljudi .
Veliko tavžent in tavžent jih časno ino večno gori ; ter se jih od leta do leta več vnema .
Ako se ne združimo tiga *kužniga* ognja pogasit , nam hoče v kratkim naš zarod *vgonobit* ' .
— *Vsmilite* se , *vsmilite* se vi , *prijateli* , ino gasiti pomagajte !
Kje pa je ogenj ?
— vprašate .
Po *tabernah* ino na *poljah* , po *zhodah* in domačijah — žganje na mizi in po omarah , ki ga ljudje ne po *glažekah* , ampak le po *poličah* če dalje bolj pijejo ; ino več ko ga , spijejo , hujši gorijo , pamet in zdravje — dušo in truplo zgubijo .
— *Povéte* , ali ni žganje iz pekla domá — očitno delo satana ?
— peklenska *pjača* , ki se naše dni po sveti vozi ino ljudi moti ?

Fig. 3 GUI for collaborative lexicon building: document mode with highlighting, in which different types of words are presented in different colours.

3.3 Adapting the system to Slovene

In addition to integrate language specific background resources, as for example, the rewrite pattern set and the modern lexicon into the system, internal software adjustments were also necessary. To start lexicon building, the tokenizer had to be enabled to cope with special characters, such as *ſ* and *Ş* introduced by the alphabet of historical Slovene. Furthermore, the software had to be adapted to the specific format of the Slovene modern lexicon that differs from the format used for the German variant of the lexicon building tool. Since the Slovene rewrite patterns include word boundaries, the contextual treatment of patterns had to be extended accordingly. During the first round of lexicon production a list of further issues turned up which is reported on in the next section.

4. Discussion

Currently only a few hundred word-forms have been added to the lexicon of historical Slovene; rather than going for quantity, we first concentrated on the various types of issues that arise in lexicon construction. In this section we discuss the main problems – and solutions – that we have encountered in our work so far. Below we provide a typology of cases which arise in the construction of the lexicon. The first five are already a part of LeXtractor itself:

1. *Historical word-forms without descendants*: such is a case with pairs of similar words, of which only one survived in modern Slovene; the other is thus included under this category. For example, both *pervle* and *pervič* (*firstly*$_{[adv.]}$) existed in 19th century, but only *prvič* is used in modern Slovene.
2. *Problematic historical word-forms*: although there is such a proposed category in LeXtractor, we are currently avoiding including words in it, as we are still determining the general methodology of how to deal with such cases.
3. *Named Entities*: when entities are pattern-based, they pose a specific problem. Such was the case with the given name *Ménart*. Since modern Slovene (usually) does not include diacritical marks in writing, patterns for diacritic removal were added, so that the word-form can be found in the modern background lexicon. The LeXtractor tool does not allow for adding the word-forms to the list of entities once a pattern has been chosen. This leads to making a list of attestations where a word-form is used as a name, adding the attestations under sub-lexicon Entities, and manually adding the modern string without the diacritical mark, in our case, *Menart*.
4. *Modern word-form missing in modern background lexicon*: some word-forms are still alive in modern (though not necessary standard) Slovene, but are missing from our contemporary background lexicon. Such is the case of the word *tavžent*, a deformed German word *tausend* (*thousand*$_{[num.]}$), that is missing from the background lexicon, even though the word is very much alive in spoken Slovene. Another example is word-form *Ogerska* (*Hungary*$_{[sg.,loc.]}$), which only exists as an adjective in the background lexicon, and not as a geographical name, as is the case in historical corpus. The proposed solution is either to add these words in modern background lexicon, or to have an edit option in LeXtractor, which would allow the lexicographer to add MSD information to the word-form.
5. *Identical word-forms*: in the otherwise trivial case when a historical word-form exactly corresponds to a modern word-form it can happen that the entry in the lexicon is a false friend; for example, *serca* is an archaic form for *srca* (*heart*$_{[sg,gen]}$), but at the same time is identical to a form of the contemporary lemma *serec* (*horse of a gray colour*). The problem is solved when we ascribe a pattern to the word-form, *r→er*, which transcribes *serca* into a univocal *srca*.
6. *Missing readings*: sometimes the correct MSD is missing in the background lexicon, and is consequently also missing in LeXtractor. Such was the case with word-forms *dobé* and *mrtve* that represent two different approaches of dealing with such problems. The word-form *dobé* was transcribed into *dobe* (both *they get* and *era*$_{[sg.\ gen.\ \text{or}\ pl.\ nom.]}$), but the background lexicon only offered the latter reading, i.e. *Ncf: noun, common, feminine*. We decided to change the pattern (*ijo→e*) and modernize the otherwise possible, but in contemporary Slovene archaic verb form *dobe* into a more common modern form *dobijo*. This extracted the proper reading for *dobé* from the background lexicon. The second possible scenario of a missing MSD is much more complicated. The possible solution for it is the same solution we propose for modern word-forms missing in modern background lexicon. Sometimes a historical word-form was used differently than modern word-form, which means that the modern background lexicon cannot offer the missing MSD. For example, the word-form *mrtve*, it is not only an adjective of a feminine plural for *dead*, it is also an accusative plural form for a masculine noun *mrtvi* (*the dead*), as well as a nominative plural form for a feminine noun *the dead*. This last use, however, is foreign to contemporary Slovene, which only uses masculine plural noun *mrtvi*. An edit option, with which the lexicographer could manually add the missing MSD information, would again be needed.
7. *Historical word-form corresponds to more than one modern word-form of the same lemma*: such is a case with the word-form *veči*, that doesn't only transcribe into *večji* (*bigger*), but also into *večja*, *večje* and *večjo* (*veči žejo*/*večjo žejo*, *veči groze*/*večje groze*, *veči del*/*večji del*, *veči nesreča*/*večja nesreča* etc.). The solution to this is to first add the suggested readings and attestations and then create additional entries manually. The lexicographer needs to pay special attention with such cases, because the entry procedure must be performed in a single step: once we stop working on the word-form *veči* and work on another word-form, the only way to return *veči* is to destroy all the entries for it, recalculate and start again. Once again an edit option would be needed.
8. *Change of inflection*: a form of the missing reading or MSD is an inaccurate reading. Sometimes the word form for a specific declination has changed during

time: for example, the word-form *serci* (from *srce, heart*), was used both as the nominative and accusative dual form, *najne serci* (*our two hearts*), as well as the historical singular dative and locative form, *k serci*, *pri serci*. Contemporary Slovene uses the form *srci* only for the first case (and as instrumental plural form, though no such attestations were found in historical corpus), whereas the modern singular dative form is *srcu*. The dilemma that arises is this: should a new pattern, and hence a new reading, be added transcribing *u→i*, so that the full and correct MSD can later be extracted not only for the dual but also for the singular dative form? If this is not done and the lexicographer just ascribes the information, that *srci* is a common neutral noun, later, more specific MSD extraction could not recognize that *srci* is also a historical singular dative form.

9. *The background lexicon offers too many possible readings*: another frequent problem that needed systematical solving is a case when a word-form, transcribed into a modern word-form, offers more grammatical readings (MSDs) than the historical corpus shows are needed. For example, the word form *ravna* only appeared in the historical corpus as a form of a verb *ravnati* (*to straighten*), even though it could also be a feminine adjective form of *raven* (*straight*). There are two possible solutions for such a case. One is to exclude the reading for adjective because the word-form historically did not exist as an adjective, the other is to mark that there were no attestations found, even though the word-form *ravna* already existed as an adjective. We have opted to exclude the reading when the word as such doesn't appear in Pleteršnik's dictionary, published in 1894—1895, and on the other hand mark that there were no attestations found for the proposed reading, if the word does appear in his dictionary, as was the case for *ravan.* In the future, when texts before 1847 are added, it would also be wise to include cited older dictionaries as a reference.
10. *Historical word-form is written separately, modern word-form is not*: sometimes, historical word-forms are written separately, even though in modern Slovene they are written as one word. Such is the case of compound words *najprej* (*firstly*), historically *nar pervo*, *zase* (*for him/her/itself*), historically *za-se*, and *čezenj* (*over him*), historically *čeznj*, *čez-nj*, and in one instance also *čes-nj*. Words with the prefix *nar* were sometimes written separately, like *nar pervo*, and sometimes together, *narbolj* (*mostly*). Since some compound words were written with a hyphen and others were not, a possible solution would be to merge the prefix with the following word, so that LeXtractor recognizes the compound word in the background lexicon, with or without applying patterns. A solution more complex to implement would be for the lexicographer having the option of marking that two words actually form one and ascribe modern compound word to the unit.

5. Conclusions

The paper presented the first steps in building a lexicon of XIXth century Slovene, using a historical corpus, a contemporary lexicon of Slovene, spelling variation patterns, and the LeXtractor software. So far we have mostly concentrated on setting up the resource and program environment and methodological issues, which have been discussed in the present paper.

In further work we plan to intensively start adding entries to the lexicon, extend the corpus, esp. with newspapers and older books, as well as address the remaining methodological issues, such as tokenisation, which, as discussed, can be different in historical words from their contemporary equivalents.

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

Acknowledgements

The work presented in this paper was supported by the EU FP7 ICT project IMPACT, "Improving Access to Text".

References

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 and Sašo Džeroski. 2004. Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18(1):17–41.

Tomaž Erjavec. 2007. Architecture for Editing Complex Digital Documents. Proceedings of the Conference on Digital Information and Heritage. Zagreb. pp. 105-114.

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

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 [German-Slovene/Croatian translation, 1848-1918]. *Ein Werkstättenbericht. Wiener Slavistisches Jahrbuch 53/2007*. Austrian Academy of Sciences Press, Vienna. pp. 163-176.