CC BY-NC-SA-4.0 | DOI: https://doi.org/10.3986/9789610508861_02

Tomaž Erjavec

The Collection of Slovenian paremiological units Pregovori: Encoding, publication and quantitative overview

Zbirka slovenskih paremioloških enot Pregovori: kodiranje, objava in kvantitativni pregled

The article discusses the digital collection of paremiological units called Pregovori (Proverbs), the first large and openly available dataset of this type for the Slovenian language. We discuss its encoding in TEI XML, including linguistic annotation, and elaborate on its availability in the repository of the Slovenian research infrastructure CLARIN.SI, along with its integration into concordancers that enable on-line analyses of the Collection. We then give a short quantitative overview of the Collection, followed by the use of its constituent paremiological units in metaFida, the largest corpus of the Slovenian language. Finally, we present our conclusions and give directions for further work.

◆ **Keywords**: proverbs, digital database, text encoding initiative, language corpus, linguistic annotation, concordancers, open access

Članek obravnava digitalno zbirko paremioloških enot Pregovori, ki je prva obsežna in javno dostopna zbirka tovrstnih podatkov za slovenski jezik. Obravnavamo njeno kodiranje v TEI XML, vključno z jezikovno anotacijo, in podrobneje predstavljamo njeno dostopnost v repozitoriju slovenske raziskovalne infrastrukture CLARIN.SI ter njeno vključitev v konkordančnike, ki omogočajo spletne analize zbirke. Nato podamo kratek kvantitativni pregled Zbirke, ki mu sledi uporaba njenih sestavnih paremioloških enot v metaFidi, največjem korpusu slovenskega jezika. Na koncu predstavimo svoje zaključke in podamo usmeritve za nadaljnje delo.

◆ **Ključne besede**: pregovori, digitalna podatkovna zbirka, pobuda za kodiranje besedil, jezikovni korpus, jezikovna anotacija, konkordančniki, odprti dostop

1 Introduction

The Open Science paradigm, which advocates for the open availability of research publications and datasets, is increasingly gaining ground, even in the humanities. In this paper, we demonstrate this principle on a large dataset of Slovenian paremiological units and show how it is structured, encoded and made available for download and on-line analysis within the framework of the Slovenian research infrastructure for language resources and technologies CLARIN.SI.

The collection of paremiological units (henceforth the Collection) was initially prepared at the Institute of Slovenian Ethnology at the Scientific and Research Centre of the Slovenian Academy of Sciences and Arts. This institute serves as an archive for ethnological materials and saw a large augmentation of paremiological units during the 1990s and later years, thanks to a series of dedicated projects (Stanonik 1996; 2004; 2009; 2015). The Collection was digitised here, first by manual input into Word and later transferred to Excel.

The process of cleaning up the Excel spreadsheets and converting them to XML has already been detailed in Babič and Erjavec (2022). Here, we only mention that the input for this conversion consisted of two TSV (tab-separated values) tabular files exported from Excel. One file contained the paremiological units and the identifiers (IDs) of the bibliographic sources they were observed in (along with the page number of the mention of the paremiological unit in the source). The second file contained the list of these bibliographic units with their respective IDs. The basic conversion to XML was then relatively straightforward, particularly after the XML encoding had been defined. The XML encoded Collection is the point of departure for this paper. It should also be mentioned that the digitised collection has been published in two versions to date: the "Collection of Slovenian paremiological units Pregovori 1.0" (Babič et al. 2022), and the "Collection of Slovenian paremiological units Pregovori 1.1" (Babič et al. 2023), which corrected some errors from version 1.0 and also slightly extended its scope. In this paper we provide an overview of version 1.1.

The rest of this paper is structured as follows: Section 2 discusses the encoding of the Collection, including its metadata, the bibliographical sources and the paremiological units, as well as the automatic linguistic analysis of the units. Section 3 overviews the publication of the Collection, detailing its availability for download from the CLARIN.SI repository and access through the CLARIN.SI concordancers. Section 4 gives a quantitative overview of the Collection, first in terms of its size and other characteristics, and then in terms of the usage of its paremiological units in a large Slovenian language corpus. Finally, Section 5 gives the conclusions and some directions for further work.

2 Encoding of the Collection

The Text Encoding Initiative (TEI) Guidelines (TEI Consortium, 2020) is a robust and comprehensive framework for encoding and analysing textual materials. TEI originated in the mid-1980s when the academic community recognised the need for a standardised approach to representing and exchanging textual information in the burgeoning field of digital humanities. At the core of the TEI mission was the development of guidelines to enable scholars, librarians and archivists to create machine-readable texts capable of capturing the richness and complexity of human expression.

The primary goal of the TEI is to provide a set of guidelines and standards for encoding texts in a way that is both human-readable and machine-readable. TEI markup, expressed in XML (eXtensible Markup Language), allows for the identification and description of various elements within a text. This granular encoding facilitates sophisticated analysis, searching and presentation of textual content, transcending the limitations of traditional print media. One of the notable features of TEI is its adaptability. The guidelines are designed to accommodate a diverse range of texts, from literary works and historical documents to linguistic corpora and multimedia resources. This flexibility has made TEI a valuable tool for scholars across different disciplines, fostering interdisciplinary research and collaboration.

TEI has had a profound impact on the field of digital humanities. Scholars and institutions around the world have adopted TEI Guidelines as a standard for encoding and exchanging textual data. This widespread acceptance has facilitated the creation of digital archives, electronic editions of texts and innovative research projects that leverage the capabilities of TEI-encoded materials. Moreover, TEI has contributed to the democratisation of access to cultural heritage. By encoding and digitising historical documents, libraries, museums and archives can make their collections available to a global audience, fostering a new era of scholarship and exploration. TEI-encoded texts also enable advanced computational analysis, opening doors to new research methodologies and insights.

For these reasons, the Collection is encoded in TEI. In particular, it is modelled as one TEI document (XML element *TEI*), which then contains the TEI header (element *teiHeader*) giving the metadata (including the bibliographical sources) of the Collection, followed by the text (element *text*), i.e. the paremiological units constituting the Collection. The next subsection explains and illustrates the encoding of the TEI header, while the following subsection delves into the encoding of paremiological units. We conclude the Section by explaining the automatic linguistic processing of the Collection and how it is encoded in TEI.

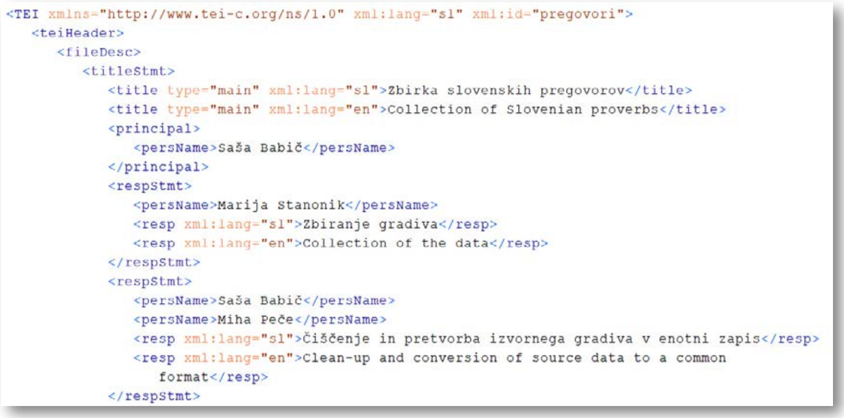
2.1 The TEI header

We illustrate the start of the document in Figure 1, giving the overall *TEI* root XML tag with the attributes for the TEI namespace, and the ID and language of the resource. The TEI header is the first element inside *TEI*, and it is illustrated next, specifically the file description (*fileDesc*), which encompasses the title statement (*titleStmt*) containing the basic bibliographic metadata of the document.

The TEI header contains other substantial metadata. Here, we highlight only two notable elements of the Collection's metadata. The first is the taxonomy of bibliographic units that were the source of the proverbs. The bibliographic units were classified into 18 types (the complete list is given in Table 1) and these types are encoded in the TEI header class declarations (*classDecl*), specifically within the *taxonomy* element. The encoding of the first six types is given in Figure 2.

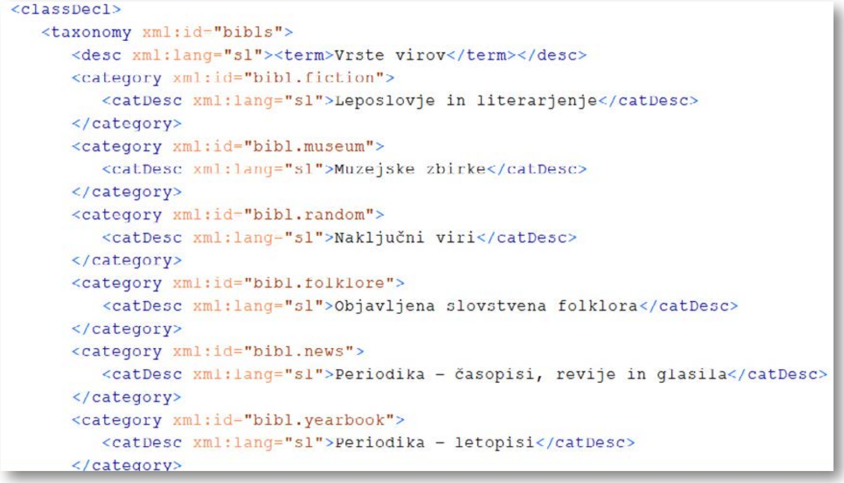
As can be seen, each category is assigned an ID and then given the Slovenian description of the category in the category description (*catDesc*) element.

The second element illustrated is the source description (*sourceDesc*), specifically its list of bibliographic sources (*listBibl*), with each source encoded in the bibliography element (*bibl*), as illustrated for a few types of sources in Figure 3.



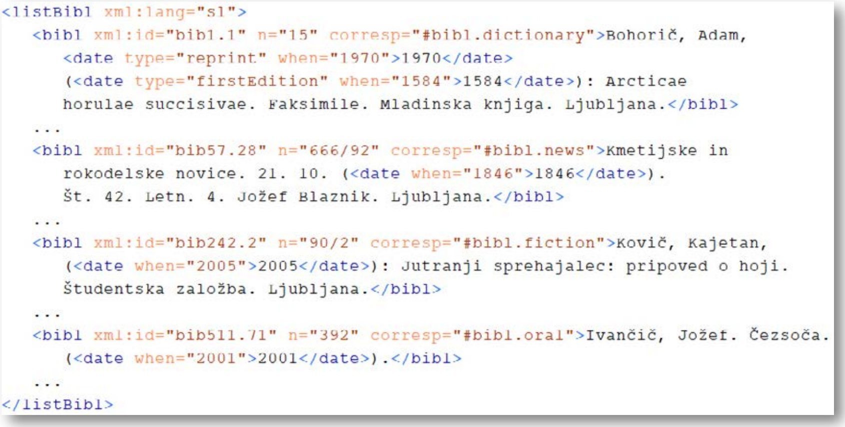
```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="sl" xml:id="pregovori">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title type="main" xml:lang="sl">Zbirka slovenskih pregovorov</title>
            <title type="main" xml:lang="en">Collection of Slovenian proverbs</title>
            <principal>
               <persName>Saša Babič</persName>
            </principal>
            <respStmt>
               <persName>Marija Stanonik</persName>
               <resp xml:lang="sl">Zbiranje gradiva</resp>
               <resp xml:lang="en">Collection of the data</resp>
            </respStmt>
            <respStmt>
               <persName>Saša Babič</persName>
               <persName>Miha Peče</persName>
               <resp xml:lang="sl">Čiščenje in pretvorba izvornega gradiva v enotni zapis</resp>
               <resp xml:lang="en">Clean-up and conversion of source data to a common
                  format</resp>
            </respStmt>
```

Figure 1: The beginning of the Collection TEI document.



```
<classDecl>
   <taxonomy xml:id="bibls">
      <desc xml:lang="sl"><term>Vrste virov</term></desc>
      <category xml:id="bibl.fiction">
         <catDesc xml:lang="sl">Leposlovje in literarjenje</catDesc>
      </category>
      <category xml:id="bibl.museum">
         <catDesc xml:lang="sl">Muzejske zbirke</catDesc>
      </category>
      <category xml:id="bibl.random">
         <catDesc xml:lang="sl">Naključni viri</catDesc>
      </category>
      <category xml:id="bibl.folklore">
         <catDesc xml:lang="sl">Objavljena slovstvena folklora</catDesc>
      </category>
      <category xml:id="bibl.news">
         <catDesc xml:lang="sl">Periodika – časopisi, revije in glasila</catDesc>
      </category>
      <category xml:id="bibl.yearbook">
         <catDesc xml:lang="sl">Periodika – letopisi</catDesc>
      </category>
```

Figure 2: The taxonomy of bibliographical types.



```
<listBibl xml:lang="sl">
   <bibl xml:id="bib1.1" n="15" corresp="#bibl.dictionary">Bohorič, Adam,
      <date type="reprint" when="1970">1970</date>
      (<date type="firstEdition" when="1584">1584</date>): Arcticae
      horulae succisivae. Faksimile. Mladinska knjiga. Ljubljana.</bibl>
   ...
   <bibl xml:id="bib57.28" n="666/92" corresp="#bibl.news">Kmetijske in
      rokodelske novice. 21. 10. (<date when="1846">1846</date>).
      Št. 42. Letn. 4. Jožef Blaznik. Ljubljana.</bibl>
   ...
   <bibl xml:id="bib242.2" n="90/2" corresp="#bibl.fiction">Kovič, Kajetan,
      (<date when="2005">2005</date>): Jutranji sprehajalec: pripoved o hoji.
      Študentska založba. Ljubljana.</bibl>
   ...
   <bibl xml:id="bib511.71" n="392" corresp="#bibl.oral">Ivančič, Jožef. Čezsoča.
      (<date when="2001">2001</date>).</bibl>
   ...
</listBibl>
```

Figure 3: Examples of bibliographic sources for the proverbs.

Each bibliographic element is assigned an ID, with the generic label (*n*) encoding the element's identifier in the physical collection at the Institute of Slovenian Ethnology. The elements also contain the correspondence (*corresp*) pointer to the categories of the previously mentioned taxonomy, effectively classifying the bibliographic source into one of the 18 defined types.

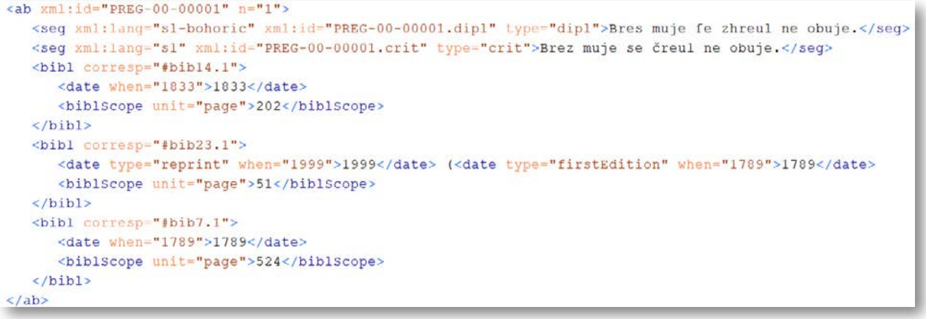
The content of the bibliographical units is semi-structured, i.e. they contain the text describing the unit, with the date(s) of publication marked up (date). The year attribute (*when*) specifies either the publication year of the source or, in the case of serial publication such as "1680–1685", denotes the start and end dates of its publications (attributes *from* and *to*). Assigning temporal information to controlled attribute values enables automatic processing. For example, it enables the selective retrieval of paremiological units gathered from sources within a specific range of years. If the source is a reprint, this date is distinguished from that of the first edition of the work by the value of the *type* attribute.

TEI allows much more detailed and structured bibliographic descriptions. However, authors and titles are given in a variety of ways, and encoding these additional distinctions would require the extensive list of sources to be marked up manually.

2.2 Markup of the proverbs

In the *text* element of the TEI document, the individual proverbs are marked up as anonymous blocks (*ab*), where each paremiological unit is presented both in the orthography of the source (referred to as the diplomatic transcription), and modernised (standardised) in terms of its alphabet (critical transcription). Each unit is accompanied by bibliographic sources, as illustrated in Figure 4.

The two transcriptions are encoded as segments (*seg*) of two types. Each segment is also marked with its language, where *sl* is the ISO two-letter code for (contemporary) Slovenian, while *sl-bohoric* is the IANA code for the Bohorič alphabet,[1] which was employed until about the 1850s.



```
<ab xml:id="PREG-00-00001" n="1">
   <seg xml:lang="sl-bohoric" xml:id="PREG-00-00001.dipl" type="dipl">Bres muje ſe zhreul ne obuje.</seg>
   <seg xml:lang="sl" xml:id="PREG-00-00001.crit" type="crit">Brez muje se čreul ne obuje.</seg>
   <bibl corresp="#bib14.1">
      <date when="1833">1833</date>
      <biblScope unit="page">202</biblScope>
   </bibl>
   <bibl corresp="#bib23.1">
      <date type="reprint" when="1999">1999</date> (<date type="firstEdition" when="1789">1789</date>
      <biblScope unit="page">51</biblScope>
   </bibl>
   <bibl corresp="#bib7.1">
      <date when="1789">1789</date>
      <biblScope unit="page">524</biblScope>
   </bibl>
</ab>
```

Figure 4: Examples of bibliographic sources for the proverbs.

[1] https://www.iana.org/assignments/lang-subtags-templates/bohoric.txt

Short editorial notes from the source materials have also been preserved in the diplomatic transcription. They usually give the time of year or date, for example, for saints' days:

(1) *Pred svetim Jakobom<note>25. julij</note> tri dni lepó, rž prav redno dozorela bo.*
("Before St. Jakob's<note>July 25th </note> three days nice, rye ripens in a trice.").

The bibliographic elements (*bibl*) point to (*corresp* attribute) the bibliographic sources of the units (*sourceDesc/listBibl/bibl*). The page(s) where the unit can be found in the source are given in the scope of bibliographic reference (*biblScope*). The date(s) of publication are also given for convenience, even though they are redundant, as they are also given in the bibliographical source in the TEI header.

2.3 Linguistic annotation

While encoded in TEI, the information presented so far is essentially the same as in the two tables of bibliographical sources and paremiological units, which served as the basis for the TEI conversion. The Collection is also distributed in this form, for those only interested in the plain text of the paremiological units. However, another variant of the Collection has also been prepared. It is largely identical to the "plain text" version but has added segments that contain the automatically linguistically annotated text of the critical transcriptions. This version, with its added linguistic annotation, enables better searching through the collection (e.g. by lemma rather than word form), as well as searching via abstract linguistic categories, such as the part-of-speech or syntactic relations.

In the first stage of the linguistic processing, the critical transcriptions of the proverbs were tokenised, i.e. split into words and punctuation symbols. In cases where the proverb comprised more than one sentence they were also segmented into sentences. The second stage involved modernising the identified words. The reason for this is twofold. Firstly, words used to be written differently in historical Slovenian, differing not only from the way they are written today but the spelling also depended on the author and the exact time period, as it had not yet been standardised. This variation means it would otherwise be much more difficult to search for specific words in the Collection, so they have been modernised and their spelling unified. Secondly, the models used for further linguistic analysis have been trained on datasets of contemporary standard Slovenian and work much better on modernised words than when applied to words with archaic spellings.

For modernising the words, we used the open-source normalisation tool[2] cSMTiser (Scherrer and Ljubešić 2016), which is based on the Statistical Machine Translation (SMT) tool Moses (Koehn 2010). However, while Moses and other SMT tools translate words in a sentence, cSMTiser translates individual letters (characters) in a word. The

[2] https://github.com/clarinsi/csmtiser

tool was trained to carry out modernisation using the manually modernised corpus of historical Slovenian goo300k (Erjavec 2015). This is a process that is similar to what was previously undertaken for modernising the words in the collection of historical Slovenian novels as part of the ELTeC corpus (Schöch et al. 2021). The tool was then used to modernise the word tokens in the critical transcriptions. It should be noted that like other tools based on machine learning principles, cSMTiser knows how to handle unknown words (i.e. words not present in the training data). However, it also makes mistakes and does not always correctly translate archaic words into their modern-day standard spelling.

Once the words had been normalised, the resulting sentences were further annotated with the open-source tool-chain CLASSLA-Stanza[3] (Ljubešić and Dobrovoljc 2019; Terčon and Ljubešić 2023), which adds the following information to each sentence token:

- The lemma or base form of the word, an essential piece of information that simplifies searching for the highly inflective Slovenian language.
- The morphosyntactic description (MSD) according to the MULTEXT-East specifications[4] (Erjavec 2012). For example, *Ncmsg* for the morphosyntactic features *Noun, Type = common, Gender = masculine, Number = singular, Case = genitive*. It should be noted that conversion tables exist for mapping from MSDs to their features, as well as to MSDs and their features in Slovenian. For example, the Slovenian equivalent of *Ncmsg* is the MSD *Somer* and the features *samostalnik, vrsta = občno_ime, spol = moški, število = ednina, sklon = rodilnik*.
- The morphological features according to the Universal Dependencies framework (de Marneffe et al. 2021)[5] for Slovenian (Dobrovoljc et al. 2017), e.g. *NOUN, Case = Gen, Gender = Masc, Number = Sing*. These features are similar to those of MULTEXT-East, however, the features have different names and there are sometimes some systematic differences. The reason why both are present is that the MULTEXT-East MSDs have a long tradition of use in Slovenian corpora, while the Universal Dependencies are now becoming a de-facto standard for many languages including Slovenian.
- The syntactic parse of the sentence, also according to the Universal Dependencies formalism. The parse is dependency-based, i.e. head/argument pairs of sentence tokens are linked and the link is associated with its dependency label.

Figure 5 illustrates the added linguistic information for one paremiological unit, which is encoded in the segment typed as being the normalised form of the critical

[3] https://github.com/clarinsi/classla

[4] https://github.com/clarinsi/mte-msd, http://nl.ijs.si/ME/V6/msd/html/msd-sl.html

[5] https://universaldependencies.org/

transcription (*type="crit.norm"*). As can be noted, all the words in this example are identical to the ones in the critical transcription, which are indeed spelled the same in the contemporary standard, except for "čevlj" as the modernised form of "čreul". This modernisation illustrates the already mentioned fact that cSMTiser also makes mistakes, as the correct form should in fact be "čevelj". Nevertheless, the automatically modernised form is closer to the standard than the one in the critical transcription.

The second added segment (only the beginning of which is shown to save space) is presented in its normalised and linguistically analysed critical transcription (*type* = *"crit.norm.ana"*). This segment contains one or more sentences (element *s*), each of which then contains tokens – either words (*w*) or punctuation symbols (*pc*, not shown here). The words themselves are presented as they appear in the critical transcription. Any variations in their normalised form are indicated by the value of the normalised form (*norm*) attribute. The MULTEXT-East MSD is given as the value of the analysis (*ana*) attribute, serving as a pointer to the decomposition of the MSD into features. The Universal Dependencies morphological features are given as the value of the MSD (*msd*) attribute, and the lemma as the value of the *lemma* attribute. Each token also has an ID, which is necessary for encoding the syntactic analysis, further explained below.

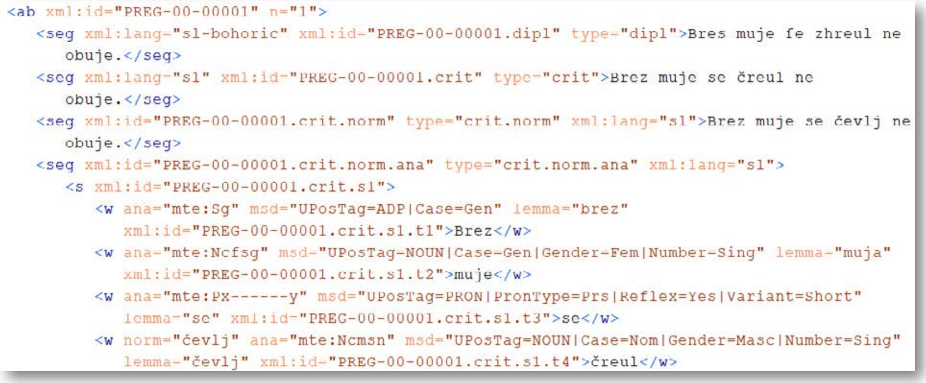
A complication arises in cases where one word is normalised into several contemporary words. For example, "nevboga" ("does not obey") is now written as "ne uboga" and the "not" is no longer a prefix but a word. The encoding of such a 1-2 mapping is illustrated in Figure 6 (the IDs of the various elements have been removed to improve readability).

The word in focus, "nevboga" is marked up using the word element (w), as usual. However, it does not have the normalisation attribute or any other linguistic analysis attributes. Instead, it contains two-word elements, which do have the normalisation attribute ("ne" + "uboga") along with all the other linguistic analysis attributes. However, these two words have no content.

The example also illustrates the use of the punctuation character (*pc*) element for the comma, as well as the use of the join attribute (*join="right"*) on the penultimate token, meaning that there is no space between "uboga"/"nevboga" and the following comma.

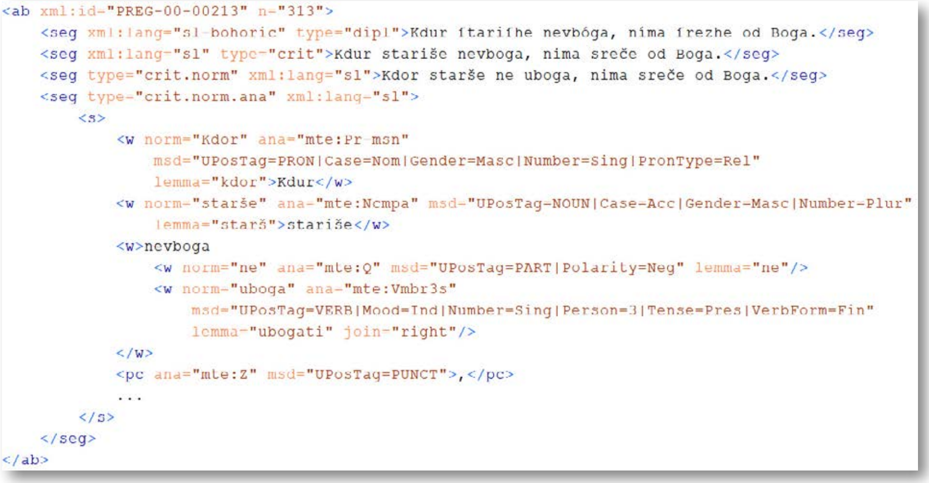
Finally, the syntactic analysis is encoded inside its sentence (*s*) element, in a standoff fashion. This means it is not encoded as attributes on the tokens, but rather in a special link group (*linkGrp*) element. Each constituent link (*link*) in this group points to the IDs of the relevant tokens, as illustrated in Figure 7. The example shows the last token of the sentence, i.e. the final period, followed immediately by the link group.

The link group is presented as encoding Universal Dependencies syntax (*type="UD-SYN"*) and its target function (*targFunc="head argument"*), which describes the function of each of the values of the *target* attribute in the contained links, stating that these are the head and argument of the syntactic relation. It also states that the syntactic parse corresponds (*corresp*) to the sentence with the ID *PREG-00-00001.crit.s1*.



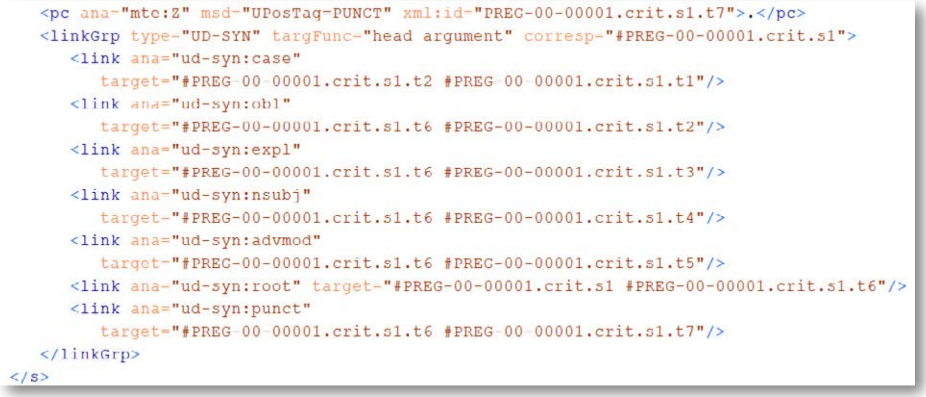
```
<ab xml:id="PREG-00-00001" n="1">
   <seg xml:lang="sl-bohoric" xml:id="PREG-00-00001.dipl" type="dipl">Bres muje ſe zhreul ne
      obuje.</seg>
   <seg xml:lang="sl" xml:id="PREG-00-00001.crit" type="crit">Brez muje se čreul ne
      obuje.</seg>
   <seg xml:id="PREG-00-00001.crit.norm" type="crit.norm" xml:lang="sl">Brez muje se čevlj ne
      obuje.</seg>
   <seg xml:id="PREG-00-00001.crit.norm.ana" type="crit.norm.ana" xml:lang="sl">
      <s xml:id="PREG-00-00001.crit.s1">
         <w ana="mte:Sg" msd="UPosTag=ADP|Case=Gen" lemma="brez"
            xml:id="PREG-00-00001.crit.s1.t1">Brez</w>
         <w ana="mte:Ncfsg" msd="UPosTag=NOUN|Case=Gen|Gender=Fem|Number=Sing" lemma="muja"
            xml:id="PREG-00-00001.crit.s1.t2">muje</w>
         <w ana="mte:Px------y" msd="UPosTag=PRON|PronType=Prs|Reflex=Yes|Variant=Short"
            lemma="se" xml:id="PREG-00-00001.crit.s1.t3">se</w>
         <w norm="čevlj" ana="mte:Ncmsn" msd="UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Sing"
            lemma="čevlj" xml:id="PREG-00-00001.crit.s1.t4">čreul</w>
```

Figure 5: Example of a linguistically annotated proverb.



```
<ab xml:id="PREG-00-00213" n="313">
    <seg xml:lang="sl-bohoric" type="dipl">Kdur ſtariſhe nevbóga, níma ſrezhe od Boga.</seg>
    <seg xml:lang="sl" type="crit">Kdur stariše nevboga, nima sreče od Boga.</seg>
    <seg type="crit.norm" xml:lang="sl">Kdor starše ne uboga, nima sreče od Boga.</seg>
    <seg type="crit.norm.ana" xml:lang="sl">
        <s>
            <w norm="Kdor" ana="mte:Pr-msn"
                msd="UPosTag=PRON|Case=Nom|Gender=Masc|Number=Sing|PronType=Rel"
                lemma="kdor">Kdur</w>
            <w norm="starše" ana="mte:Ncmpa" msd="UPosTag=NOUN|Case=Acc|Gender=Masc|Number=Plur"
                lemma="starš">stariše</w>
            <w>nevboga
                <w norm="ne" ana="mte:Q" msd="UPosTag=PART|Polarity=Neg" lemma="ne"/>
                <w norm="uboga" ana="mte:Vmbr3s"
                    msd="UPosTag=VERB|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin"
                    lemma="ubogati" join="right"/>
            </w>
            <pc ana="mte:Z" msd="UPosTag=PUNCT">,</pc>
            ...
        </s>
    </seg>
</ab>
```

Figure 6: Example of a 1-2 mapping between the diplomatic and critical words.



```
   <pc ana="mte:Z" msd="UPosTag=PUNCT" xml:id="PREG-00-00001.crit.s1.t7">.</pc>
   <linkGrp type="UD-SYN" targFunc="head argument" corresp="#PREG-00-00001.crit.s1">
      <link ana="ud-syn:case"
         target="#PREG-00-00001.crit.s1.t2 #PREG-00-00001.crit.s1.t1"/>
      <link ana="ud-syn:obl"
         target="#PREG-00-00001.crit.s1.t6 #PREG-00-00001.crit.s1.t2"/>
      <link ana="ud-syn:expl"
         target="#PREG-00-00001.crit.s1.t6 #PREG-00-00001.crit.s1.t3"/>
      <link ana="ud-syn:nsubj"
         target="#PREG-00-00001.crit.s1.t6 #PREG-00-00001.crit.s1.t4"/>
      <link ana="ud-syn:advmod"
         target="#PREG-00-00001.crit.s1.t6 #PREG-00-00001.crit.s1.t5"/>
      <link ana="ud-syn:root" target="#PREG-00-00001.crit.s1 #PREG-00-00001.crit.s1.t6"/>
      <link ana="ud-syn:punct"
         target="#PREG-00-00001.crit.s1.t6 #PREG-00-00001.crit.s1.t7"/>
   </linkGrp>
</s>
```

Figure 7: Example of a dependency parse of a proverb.

Each of the link elements in the link group then encodes a dependency syntactic relation between two tokens. The target attribute contains a pair of pointers, the first to the head and the second to the argument of the relation, while the syntactic dependency type is encoded in the linguistic analysis (*ana*) attribute. As with the MULTEXT-East MSDs for tokens, the value of this attribute serves as a pointer to the taxonomy of the Universal Dependencies syntactic relations.

It should be noted that the TEI header of the linguistically analysed Collection has a few elements not present in the plain-text version. These include the taxonomy of the Universal Dependencies syntactic relations, the definitions of extended pointers (element *prefixDef*), such as the *ud-syn* prefix in the syntactic analysis above, and the application information (*appInfo*) detailing which tools have been used to perform the linguistic analysis.

3 Publishing the Collection

Our aim was to publish the Collection in open access, with minimal barriers to its use by other researchers and interested parties, thereby contributing to the principles of Open Science. Furthermore, we also wanted to make the Collection immediately available for analysis, rather than simply offering its download. In this Section, we discuss these two methods of distribution.

3.1 The downloadable dataset

We published the Collection in the repository of the Slovenian node of the research infrastructure for language resources and technologies CLARIN.SI (Babič et al. 2023), where it is available for download under the Creative Commons – Attribution (CC BY) licence. The repository offers long-term access to the data, gives it a permanent identifier, as well as clearly stating how the dataset must be cited when used in further research, giving appropriate credit to the authors.

The Collection is available in three formats. First, there is the canonical TEI encoding, distinguishing the "plain text" variant from the linguistically analysed one. The reason for distinguishing between the two is that not everybody will be interested in the linguistic analyses, and the linguistically analysed variant is much larger than the plain-text one. The TEI encoded Collection comprises the following files:

- *pregovori-viri.xml*: the file with the bibliographic sources;
- *pregovori.xml*: the root file (i.e. the *TEI* element) of the plain-text version with the TEI header, which, using the XInclude mechanism, includes the files *pregovori-viri.xml* and *pregovori-text.xml*;
- *pregovori-text.xml*: the file with all the paremiological units (i.e. the *text* element) in their plain-text version;

- *pregovori.ana.xml*: the root file (i.e. the *TEI* element) of the linguistically annotated version with the TEI header; through the XInclude mechanism, it includes the files *pregovori-viri.xml* and *pregovori-text.ana.xml*;
- *pregovori-text.ana.xml*: the file with all the paremiological units (i.e. the *text* element), including their linguistically annotated version.

Using the TEI-encoded variant requires some programming skills and familiarity with XML and TEI. Therefore, we have converted the TEI into a simpler format, which is also made available in the repository entry. This variant is similar to the one used as the source for the TEI and consists of two files, both formatted as TSV tables with a header row and TAB-separated columns:

- *pregovori-viri.tsv*: the file with the bibliographic sources, giving the ID for each source, the ID as used in the source collection, the category of the unit, its year(s) of publication, and the name of the bibliographical unit;
- *pregovori.tsv*: the file with the proverbs, giving each ID, the source ID, the diplomatic, critical, and normalised transcription, as well as flags specifying if there is a difference between the diplomatic and critical transcription (flag *DC*) and/or between the critical and normalised transcription (flag *CN*).

Finally, the third format in which the collection is available is the so-called vertical format, automatically converted from the linguistically annotated TEI version of the corpus. Vertical files are used by many concordancers, particularly those offered by CLARIN.SI, as detailed in the next subsection. A vertical file contains XML-like tags (structural annotations), possibly with attributes, and one line for each token, giving the token itself in the first column followed by an arbitrary number of columns with further information on the token, the so-called positional attributes. It should be noted that the vertical file does not contain all the information from the source TEI, and that it uses different element and attribute names from the TEI source in order to make it more intuitive when used in concordancers. This variant in the repository entry consists of two files:

- *pregovori.vert*: the vertical file, which has the structure *text* delimiting one paremiological unit and contains attributes for the bibliographical sources, their types and years of publications, as well as an attribute for the diplomatic transcription of the proverb. The subordinate structure is the sentence (*s*), which in turn encompasses the tokens. These tokens correspond to those found in the critical transcription, featuring positional attributes such as the normalised form of the word, its lemma, the MULTEXT-East MSD in Slovenian and English, the Universal Dependencies part-of-speech, the Universal Dependencies morphological features, the ID of the token, and its syntactic dependency label to the head. The syntactic head of the token is provided with the same attributes as the token. This facilitates searching for the various properties of the syntactic head associated with each token in a concordancer.

- *pregovori.regi*: the so-called registry file, which gives the names and properties of the structural annotations and of the positional attributes, and is used by the concordancers to compile and present the corpus. This registry file is identical to the one used for the CLARIN.SI concordancers, meaning that installing the corpus on some other concordancer necessitates changing some of its values, e.g. the location of the files on the system.

3.2 The Collection in concordancers

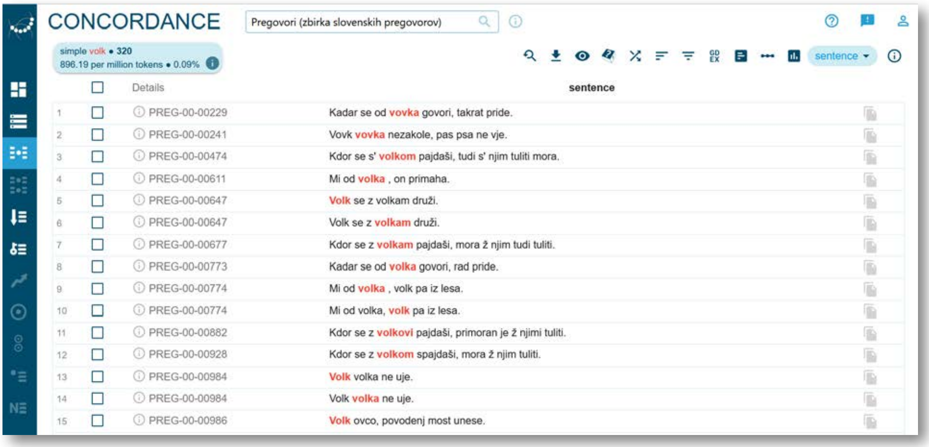
To make it possible for the Collection to be searched and analysed, we have also made it available on the CLARIN.SI concordancers, which are online tools that enable users to search text corpora and display the results in various ways. CLARIN.SI hosts two concordancers, namely noSketch Engine (Rychlý 2007; Kilgarriff et al. 2014) and KonText (Machálek 2020). While they differ in their front-end interfaces and login options, they use the same back-end – the part of the software responsible for conducting corpus searches and defining the query language for searching. This query language, called CQL (Corpus Query Language), is very powerful, as it allows searches across combinations of positional attributes and limiting searches to subcorpora defined by their structures. Furthermore, it supports searches not only by literals but also by regular expressions. For example, users can search for words (or lemmas, part-of-speech tags, etc.) that start, end, or contain a certain substring.

To give an impression of a search, Figure 8 illustrates the results of a simple search for the word “volk” (“wolf”) in the Collection. The results show 320 occurrences of “volk” in the corpus, providing the IDs of the proverbs where it appears and the proverbs themselves.

Although the object of the search was “volk”, the displayed proverbs feature this word not only in its various inflectional forms but also with archaic spelling, such as “vovka”, now written as “volka”. This occurs because simple searches not only scan the word tokens but also the normalised forms and especially the word lemmas. Such a search demonstrates the advantages of the added linguistic annotation. It should also be noted that clicking on the displayed metadata of a hit (i.e. the proverb ID) triggers a window to pop up, giving all the available metadata for a proverb.

The concordancers also support other types of analyses, with the main ones being:

- frequency lists of the chosen positional attribute;
- keywords computed against a chosen (usually reference) corpus or of a subcorpus against the whole corpus;
- collocations of the searched-for word or phrase;
- text-type analysis for the whole corpus.



CONCORDANCE

Pregovori (zbirka slovenskih pregovorov)

simple volk • 320
896.19 per million tokens • 0.09%

	□	Details	sentence
1	□	PREG-00-00229	Kadar se od vovka govori, takrat pride.
2	□	PREG-00-00241	Vovk vovka nezakole, pas psa ne vje.
3	□	PREG-00-00474	Kdor se s' volkom pajdaši, tudi s' njim tuliti mora.
4	□	PREG-00-00611	Mi od volka , on primaha.
5	□	PREG-00-00647	Volk se z volkam druži.
6	□	PREG-00-00647	Volk se z volkam druži.
7	□	PREG-00-00677	Kdor se z volkam pajdaši, mora ž njim tudi tuliti.
8	□	PREG-00-00773	Kadar se od volka govori, rad pride.
9	□	PREG-00-00774	Mi od volka , volk pa iz lesa.
10	□	PREG-00-00774	Mi od volka, volk pa iz lesa.
11	□	PREG-00-00882	Kdor se z volkovi pajdaši, primoran je ž njimi tuliti.
12	□	PREG-00-00928	Kdor se z volkom spajdaši, mora ž njim tuliti.
13	□	PREG-00-00984	Volk volka ne uje.
14	□	PREG-00-00984	Volk volka ne uje.
15	□	PREG-00-00986	Volk ovco, povodenj most unese.

Figure 8: Example of a search in the noSketch Engine.

The interface to the concordancers is RESTful, so when a query is submitted to a concordancer, the URL of the result contains all the information necessary to re-run the query. In other words, it is possible to share the URLs for particular queries with others, and the concordancers can also be queried by programs if appropriate URLs are constructed.

4 Quantitative overview

In this section, we briefly review some of the Collection's more distinctive quantitative characteristics. We first consider its size and scope, and then the use of its paremiological units in a large Slovenian language corpus.

4.1 The size and scope of the Collection

The main distinguishing feature of the Collection is its large size: it contains 37,390 paremiological units, or 286,798 words in the critical transcription. The average length of a unit is therefore 7.67 words. Of all the paremiological units, 4,556 units (12.2%) have a critical transcription different from the diplomatic one, while 8,689 units (23.2%) have an automatically normalised transcription different from the critical one. In terms of the vocabulary used, there are 33,184 different word forms that appear in the critical transcriptions of the units (29,494 if we disregard the casing), 29,615 normalised forms (26,274 disregarding casing), and 15,390 different lemmas (15,170 disregarding casing).

The list of bibliographical sources comprises 2,631 units, although only 2,599 appear as sources for the paremiological units. In other words, 32 units are listed but not used. Out of all the sources, 56 (2%) of them do not have a date of publication. The oldest source was published in 1592, three are from the 17th century, 24 (1%) from the 18th century, 667 (25%) from the 19th, and 1,880 (72%) from the 20th century. The most recent one is from 2019.

Table 1 gives the use of sources by the paremiological units categorised by their type. The first column gives the number of mentions of the source type in the Collection, the second column indicates the number of bibliographical sources corresponding to a particular type, and the third column specifies the type.

Sources	*Units*	*Type of bibliographical unit*
14,962	10	Collections of proverbs
8,067	1,124	Periodicals – newspapers, magazines and newsletters
4,940	93	Professional sources
4,080	27	Occasional sources – private collections
4,008	81	Occasional sources – collection campaigns
3,678	52	Manuscript legacies
2,989	79	Grammars and dictionaries
2,812	166	Periodicals – calendars
2,551	343	Oral sources
2,421	478	Fiction
1,168	10	Seminar and diploma theses
716	51	Museum collections
647	24	Published literary folklore
527	33	Older book sources
370	23	Random sources
214	15	Occasional sources – elementary school work
123	10	Periodicals – chronicles
11	11	Radio and television
54,284	2,630	18

Table 1: Distribution of bibliographical sources by type.

By far the largest number of mentions is from the ten printed collections of paremiological units. In this group, the largest number of mentions (4,884) is associated with the book by Etbin Bojc (1974) “Pregovori in reki na Slovenskem” (“Proverbs and Sayings in Slovenia”), which is also the overall highest-mentioned bibliographical source. As discussed in Babič and Erjavec (2022), it should be noted that this book took many of its proverbs from existing collections and older grammars and dictionaries. However, this collection is the first contemporary one.

4.2 Usage of the paremiological units in metaFida

Investigations into the use of paremiological units in corpora have a long tradition (e.g. Steyer 2017), including in the Slovenian language (e.g. Meterc 2019, 2021; Babič and Erjavec 2022). However, the method used to investigate their behaviour has typically been to concentrate on particular paremiological units (including variations). In contrast, our aim was to investigate how many of the paremiological units from the Collection are used in Slovenian and how frequently. In order to achieve this, we adopted a corpus-based approach, but rather than hand-constructing queries we wrote a script to query all the units on the CLARIN.SI noSketch Engine concordancer and simply counted the number of hits returned by the concordancer.

The corpus we queried is called metaFida (Erjavec 2023) and with 4.7 billion words, it is currently the largest corpus of Slovenian texts. It is composed of 34 individual corpora and is therefore a second-order corpus composed of previously compiled Slovenian corpora. The included corpora contain a wide variety of text types, from historical Slovene texts (the oldest one dating from 1584) to tweets and other user-generated content (the most recent texts being from 2022). The corpus was deduplicated on the paragraph level, which meant 11% of the paragraphs and 7% of the texts were removed. The annotations comprise the structures for individual texts (with attributes for the text metadata), paragraphs, sentences and gaps (for removed paragraphs). Tokens have positional attributes for the normalised form of the word (typically identical to the word, except for historical and user-generated corpora, where the words have been automatically modernised/standardised), the lemma and the MULTEXT-East MSD in Slovenian and English. The word, the normalised word and lemma attributes also have equivalent attributes for querying the lower-cased version of the attribute, e.g. *norm_lc* for a lower-cased normalised word. This makes it possible to query a particular attribute, regardless of its capitalisation in the corpus.

The script that queries the corpora takes each token of the normalised paremiological unit and converts it into a query of that token. As the paremiological units in the corpus often lack or have slightly different punctuation from those in the Collection, the punctuation tokens are queried with the MSD for punctuation (*Z* in the English MSDs), and are given as optional, i.e. with a question mark after the token. For example, if the normalised form of a unit is "*Dolgi lasje, kratka pamet.*" ("Long hair, short mind."), the constructed query is *[norm_lc="Dolgi"] [norm_lc="lasje"] [tag_en="Z"]? [norm_lc="kratka"] [norm_lc="pamet"]*. Note also that the final full-stop is omitted, as making it optional would give two hits for each attestation where the full-stop is actually present in the corpus: one hit for the unit with the full-stop and one without.

The querying of metaFida for all the paremiological units in the Collection returned 5,733 hits. This means that out of the 37,390 units, 6.5% were actually found in the corpus, indicating how many of the units in the Collection are actually used. As would be expected, the distribution is Zipfian, i.e. a few units have relatively high frequency, which then drops rapidly with roughly half being hapax legomena. More precisely:

- 2,953 (47%) of the matched units have a frequency of 1, i.e. only a single instance was found in the corpus;
- 896 (16%) of units have a frequency of 10 or more, which is at the lower limit of when certain conclusions on their use can be drawn from the context of the found units;
- 69 (1.2%) have a frequency of over 100, which is frequent enough to enable statistical analysis of the results;
- only 3 units have a frequency of over 1,000, namely:
  - 1,644: "Enkrat za vselej." ("Once and for all.")
  - 1,739: "Vaja dela mojstra." ("Practice makes perfect.")
  - 3,091: "Vsak po svoje." ("To each his own.")

It should also be noted that a fair number of hits in the metaFida corpus are of paremiological units that are mentioned rather than used in the texts. A large part of the metaFida corpus comes from the "Corpus of scientific texts from the Open Science Slovenia portal OSS" (Žagar et al. 2023), which contains 2.6 billion words from texts such as senior, masters and doctoral theses, some of which deal with proverbs and other paremiological units. Here the proverbs are not used in free text, but rather mentioned as examples, making studies of the use of such cases less than useful.

5 Conclusions

Digitising folklore materials facilitates their analysis and enables new, more exact and statistics-based methods for their study. Open access to such materials and using proven and widely used ways to encode them also means that other researchers can perform their own studies and even correct errors or add further annotations, thus increasing their value. This paper has discussed the encoding, publication and size of a large collection of Slovenian paremiological units, as well as their usage in a large Slovenian language corpus. Hopefully, all this will lead to new research into these types of short texts and enable humanities researchers in Slovenia to make further progress in the new field of Digital Humanities.

Regarding further work, the Collection almost certainly contains some typos or inconsistencies, which should be fixed. Another labour-intensive upgrade would be the manual correction of the currently automatically assigned normalised variant of the units. Alternatively, the accuracy of the automatic methods could be increased with a relatively small manually normalised sample from the Collection, and cSMTiser could also be retrained using this newly available training set.

In future, it would be interesting to explore the variation of "the same" paremiological units, adding the annotation that would selectively group them. Some units only differ in terms of punctuation or spelling, and determining which units are variants could be relatively simple to achieve with mostly automatic methods. However, a more ambitious undertaking would also consider variants that differ more substantially, e.g. using (near) synonyms, adding or removing phrases, etc.

Acknowledgements
The author would like to thank the anonymous reviewer for their very valuable comments and suggestions. The work presented in this paper was supported by the basic research project *Traditional paremiological units in dialogue with contemporary use* (ARRS J6-2579) and by the CLARIN.SI research infrastructure.

References

Babič, Saša; Erjavec, Tomaž, 2022: Izdelava in analiza digitalizirane zbirke paremioloških enot [The compilation and analysis of the digitised collection of paremiological units]. In: *Proceedings of the Conference on Language Technologies & Digital Humanities 2022*. September 15-16 2022, Ljubljana. Online: https://nl.ijs.si/jtdh22/pdf/JTDH2022_Babic_Erjavec_Izdelava-in-analiza-digitalizirane-zbirke-paremioloskih-enot.pdf.

Babič, Saša; Peče, Miha; Erjavec, Tomaž; Ivančič Kutin, Barbara; Šrimpf Vendramin, Katarina; Kropej Telban, Monika; Jakop, Nataša; Stanonik, Marija, 2022: *Collection of Slovenian paremiological units Pregovori 1.0*. Slovenian language resource repository CLARIN.SI. Online: http://hdl.handle.net/11356/1455.

Babič, Saša; Peče, Miha; Erjavec, Tomaž; Ivančič Kutin, Barbara; Šrimpf Vendramin, Katarina; Kropej Telban, Monika; Jakop, Nataša; Stanonik, Marija, 2023: *Collection of Slovenian paremiological units Pregovori 1.1*. Slovenian language resource repository CLARIN.SI. Online: http://hdl.handle.net/11356/1853.

Dobrovoljc, Kaja; Erjavec, Tomaž; Krek, Simon, 2017: The Universal Dependencies Treebank for Slovenian. In: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, 33–38. DOI: http://dx.doi.org/10.18653/v1/W17-1406.

Erjavec, Tomaž, 2012; MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation* 46/1, 35–57.

Erjavec, Tomaž, 2015; *Reference corpus of historical Slovene goo300k 1.2.* Slovenian language resource repository CLARIN.SI. Online: http://hdl.handle.net/11356/1025.

Erjavec, Tomaž, 2023: *Corpus of combined Slovenian corpora metaFida 1.0*. Slovenian language resource repository CLARIN.SI. Online: http://hdl.handle.net/11356/1775.

Kilgarriff, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vít, 2014: The Sketch Engine: ten years on. *Lexicography* 1/1, 7–36. DOI: https://doi.org/10.1007/s40607-014-0009-9.

Koehn, Philipp, 2010: *Statistical Machine Translation.* Cambridge University Press.

Ljubešić, Nikola; Dobrovoljc, Kaja, 2019: What Does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In:

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, Association for Computational Linguistics, 29–34, DOI: http://dx.doi.org/10.18653/v1/W19-3704.

Machálek, Tomáš, 2020: KonText: Advanced and Flexible Corpus Query Interface. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7003–7008. Online: https://aclanthology.org/2020.lrec-1.865.pdf.

de Marneffe, Marie-Catherine; Manning, Christopher; Nivre, Joakim; Zeman, Daniel, 2021: Universal Dependencies. *Computational Linguistics* 47/2, 255–308.

Meterc, Matej, 2019: Analiza rabe paremioloških uvajalnih sredstev v parlamentarnih razpravah iz korpusa siParl. In: Tivadar, Hotimir (ed.). *Slovenski javni govor in jezikovno-kulturna (samo)zavest*. Ljubljana: Znanstvena založba Filozofske fakultete (Obdobja, 38). 135–141.

Meterc, Matej, 2021: Aktualna raba in pomenska določljivost 200 pregovorov in sorodnih paremioloških izrazov. *Jezikoslovni zapiski* 27/1, 45–61.

Rychlý, Pavel, 2007: Manatee/Bonito - A Modular Corpus Manager. In: *Proceeding of the Conference "Recent Advances in Slavonic Natural Language Processing" (RASLAN)*, 65–70.

Scherrer, Yves; Ljubešić, Nikola, 2016: Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-Level Machine Translation. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 248–55.

Schöch, Christoph; Patraş, Roxana; Erjavec, Tomaž; Santos, Diana, 2021: Creating the European Literary Text Collection (ELTeC). *Modern languages open*. DOI: https://doi.org/10.3828/mlo.v0i0.364.

Stanonik, Marija, 1996: *Slovenski pregovori in rekla* [Slovenian proverbs and sayings]. Project proposal.

Stanonik, Marija, 2004: *Informatizacija neoprijemljive dediščine za etnologijo in folkloristiko* [The informatisation of intangible cultural heritage for ethnology and folkloristics]. Project proposal.

Stanonik, Marija, 2009: *Slovenski pregovori kot kulturna dediščina: klasifikacija in redakcija korpusa* [Slovenian proverbs as cultural heritage: classification and redaction of the corpus]. Project proposal.

Stanonik, Marija, 2015: Slovenski pregovori kot kulturna dediščina. Klasifikacija in redakcija korpusa [Slovenian proverbs as cultural heritage: Classification and redaction of the corpus]. *Traditiones* 44/3, 171–214.

Steyer, Kathrin, 2017: Corpus Linguistic Exploration of Modern Proverb Use and Proverb Patterns. In: Mitkov, Ruslan (ed.), *EUROPHRAS 2017. Computational and corpus-based phraseology: Recent advances and interdisciplinary approaches. Proceedings of the Conference Volume II (short papers, posters and student workshop papers)*. London, Geneva: Editions Tradulex, 45–52.

TEI Consortium, 2022: *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Online: https://tei-c.org/guidelines/P5.

Terčon, Luka; Ljubešić, Nikola, 2023: *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages*. arXiv 2308.04255. DOI: https://doi.org/10.48550/arXiv.2308.04255.

Žagar, Kristjan; Ferme, Marko; Ojsteršek, Milan; Jemec Tomazin, Mateja; Erjavec, Tomaž, 2023: *Corpus of scientific texts from the Open Science Slovenia portal OSS 1.0*. Slovenian language resource repository CLARIN.SI. Online: http://hdl.handle.net/11356/1774.