Progress of the RETROGRAM Project: Developing a TEI-like Model for Croatian Grammars Books before Illyrism

Petra Bago

Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences,
University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
pbago@ffzg.hr

1. Background

RETROGRAM[1] (*Retro-digitization and Interpretation of Croatian Grammar Books before Illyrism*) is a 4-year research project that started in November 2019, co-funded by the Croatian Science Foundation (IP-2018-01-3585) and the Institute of Croatian Language and Linguistics. It is a linguistic heritage project that focuses on the digitization and interpretation of pre-Illyrian Croatian grammar books with the aim to serve as a repository of such works in the future as well as to offer a model and develop processes for future similar research on digitization of Croatian grammars. So far, no digitization projects have included Croatian grammar books from the pre-Illyrian period of the Croatian language i.e. before the establishment of the common standard language[2] and orthography (Horvat and Kramarić, 2021).

Croatian language comprises of a common standard language as well as its three dialects: Čakavian, Kajkavian and Štokavian. The standardization of the Croatian literary language and the orthography based on the Štokavian dialect variant began in the 17th century. The process was finalized in the 19th century during the time of Croatian National Revival or the Illyrian movement (i.e. Illyrism). The main goals of the movement regarding language was to introduce a common literary language and a spelling reform, as well as introducing the Štokavian dialect as a linguistic common standard in order to strengthen the national cultural identity. The grammars described in this article thus belong to the pre-Illyrian period of the Croatian language, containing Croatian literary languages that precede the modern Croatian common standard language. The first grammar books were written within the religious orders, of the Jesuits and Franciscans, and were used to teach Croatian or Latin language to the Franciscan and Jesuit youth. (Horvat and Kramarić, 2021)

The main goal of the project is to create a web portal of pre-Illyrian Croatian grammar books, which would include facsimiles of selected grammar books with basic bibliographic and processing information, transcription or translation, and an index of historical grammar and linguistic terminology. The portal will be equipped with thematic searching possibilities on the morphology level. The user will be able to browse grammar books facsimiles, read transcribed or translated text, and search it by predetermined parameters (which will allow conjugation and declension paradigms search). Links to facsimiles will enable comprehensive research on orthography and traductological aspects of the selected texts. An open-access portal will be developed and available to scholars and the general public.

The main objective of the project is to intensify research activities and the interpretation of the Croatian pre-Illyrian grammars within the scope of modern linguistic disciplines (e.g. cognitive approach), to complete existing knowledge about the morphological development of the Croatian language, its normative descriptions, and development of linguistic terminology in the pre-Illyrian period. Conclusions on the formation of the Croatian language grammar model will also be based on the analysis of the Latin language grammar structure. Contrastive analysis of Latin and Croatian grammar meta-text and terminology will lead to conclusions about the influence of Latin language description on Croatian linguistic concepts in the pre-Illyrian period. More on the project can be found in Horvat (2020) and Horvat and Kramarić (2021).

2. Dataset

RETROGRAM has selected eight Croatian grammar books for the digitization and enrichment process that span from the early 17th until the early 19th century. The grammar books cover two dialects (Štokavian and Kajkavian) of pre-Illyrian Croatian before there was an agreement on the common standard language and orthography. Even though not all are grammars of Croatian language, all contain Croatian as metalanguage and/or Croatian examples of morphological paradigms. The texts are transcriptions or translations of the originals in MS Word format, as all have been published as reference books by philologists from the project's research group.

[1] https://retrogram.jezik.hr/

[2] By "common standard language" we mean a standard language covering the entire Croatian speaking area.

The selected transcriptions or translations of grammar books used for the development of the annotation model are based on the following works:

- Bartol Kašić, *Institutionum linguae Illyricae libri duo*, Rome, 1604 (Kašić, 2002),
- Jakov Mikalja, *Gramatika talijanska ukratko ili kratak nauk za naučiti latinski jezik*, Loreto, 1649 (Mikalja, Horvat, and Gabrić-Bagarić, 2008),
- Ardelio Della Bella, *Istruzioni grammaticali della lingua illirica*, Venice, 1728 (Della Bella, Sironić-Bonefačić, and Gabrić-Bagarić, 2006),
- Blaž Tadijanović, *Svašta po malo iliti kratko složenje imena, riči u ilirski i njemački jezik*, Magdeburg, 1761 (Horvat and Ramadanović, 2012),
- Marijan Lanosović, *Uvod u latinsko riči slaganje s nikima nimačkog jezika biližkama za korist slovinskih mladića složen*, Osijek, 1776 (Perić Gavrančić, 2020),
- Ignacije Szentmártony, *Einleintung zur kroatischen Sprachlehre für Deutsche*, Varaždin, 1783 (Szentmártony, 2014),
- Josip Voltić, *Grammatica illirica*, Vienna, 1803 (Voltić, 2016),
- Francesco M. Appendini, *Grammatica della lingua Illirica*, Dubrovnik, 1808 (Appendini and Lovrić Jović, 2022).

3. Data Annotation Model

The eight selected Croatian grammar books are the basis for the development of the annotation model based on the *TEI Guidelines* (TEI Consortium, 2021b). The model addresses two annotation tasks: 1) annotation of historical grammar and linguistic terminology, and 2) the annotation of morphological paradigms. The annotation tasks will be performed manually by experts working on the project. The decision was made to keep the original text intact, and any enrichment to be done through elements and attributes. Each grammar book is a TEI document comprised of a header and the body of the grammar text. The header contains metadata relevant to the project and to the particular grammar book, such as a list of all annotated grammatical terms. The body of the TEI document contains all grammar text with grammatical terminology and morphological paradigms annotations.

3.1. Grammatical Terminology Model

One of the aims of the RETROGRAM project is to facilitate research into historical grammar and linguistic terminology via the web portal. We composed an index of contemporary Croatian terms to be used for normalization of the terminology. These terms are also used in the morphological paradigms annotation task. We have identified 87 terms related to the inflected parts-of-speech. The list of terms is encoded in the TEI header. In the Example 1. we present the encoding of the term "noun"(*imenica* in Croatian) in the index to be used in the annotation model. The example is extracted from Mikalja's grammar book.

```
<encodingDesc>
 <classDecl>
  <taxonomy>
   <category xml:id="imenica">
    <catDesc>imenica</catDesc>
   </category>
    ...
  </taxonomy>
 </classDecl>
<encodingDesc>
```

Example 1: Encoding of the term "noun" (*imenica* in Croatian) in the index of Mikalja's grammar.

To annotate the term in the grammar text, we use the element `<term>`[3] that is, according to the *TEI Guidelines*, used to encode a technical term. In the Example 2 you can find encoding of the historical grammar term *IMENA* that Mikalja used to describe nouns and adjectives, hence two attribute values. The model developed for annotating grammar terminology adheres to the *TEI Guidelines*.

[3] https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-term.html

```
<p>OD <term corresp="#imenica #pridjev">IMENA</term></p>
```

Example 2: Encoding of the term "noun" (imenica in Croatian) in the grammar text of Mikalja's grammar.

3.2. Morphological Paradigms Model

For the development of the morphological paradigms model, we analyzed the following inflected parts-of-speech: nouns, pronouns, adjectives, numbers and verbs. In the *TEI Guidelines*, there is no specific module for encoding grammar texts. However, we have decided to customize the dictionary module (TEI Consortium, 2021a) since it already contains elements that group morphosyntactic information of a lexical item. Interestingly, we were not the only ones with the same idea, as Toma Tasovac and Laurent Romary addressed the issue as part of the TEI Lex-0 initiative[4]. Often the morphological paradigms are presented in a table format. For the purposes of the RETROGRAM project, we decided to disregard the presentation mode of the paradigm, and encode only the implicit information contained in the tables.

To encode one lexical item in a paradigm, we use the element `<form>`[5], which usually "groups all the information on the written and spoken form of one headword" in a dictionary. According to the *TEI Guidelines*, the element is allowed to be contained by elements grouping information on one or more entries. We violate the guidelines by allowing this element to occur in a paragraph. Except for the violation of the guidelines regarding where the element `<form>` can occur, all other child elements adhere to the TEI documentation albeit are not encoding information on a headword, but on a lexical unit of a morphological paradigm. We have defined mandatory and optional information for each inflectional parts-of-speech to be annotated as part of the RETROGRAM project, and developed a customized TEI schema. In Example 3. an encoding of two cases of the noun *vojnik* (soldier in English) as part of the paradigm is presented.

```
<p>Kad ga imenujemo, rečemo
 <form type="inflectedForm" xml:lang="hr">
  <orth>vojnik</orth>
  <gramGrp>
    <gram type="pos" corresp="#imenica"/>
    <gram type="nounType" corresp="#I_opca"/>
    <gram type="gender" corresp="#muski"/>
    <gram type="number" corresp="#jednina"/>
    <gram type="case" corresp="#nominativ"/>
    <gram type="inflectionType" corresp="#I_a_sklonidba"/>
    <gram type="animacy" corresp="#I_zivo"/>
  </gramGrp>
 </form>
 il soldato</p>
<p>Kad se pita čigovo je, rečemo
 <form type="inflectedForm" xml:lang="hr">
  <orth>vojníka</orth>
  <gramGrp>
   <gram type="pos" corresp="#imenica"/>
   <gram type="nounType" corresp="#I_opca"/>
   <gram type="gender" corresp="#muski"/>
   <gram type="number" corresp="#jednina"/>
   <gram type="case" corresp="#genitiv"/>
   <gram type="inflectionType" corresp="#I_a_sklonidba"/>
   <gram type="animacy" corresp="#I_zivo"/>
  </gramGrp>
 </form>
 del soldato</p>
...
```

Example 3: Encoding of two cases of the noun *vojnik* as segment of a morphological paradigm in Mikalja's grammar.

[4] https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Resources/grammars-in-TEI
[5] https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-form.html

4. Future Plans and Conclusion

We are currently conducting the manual annotation tasks based on the two models. Once the annotation tasks are complete, the next step is to create a web portal where all enriched grammar texts will be open and freely available with various search options.

In this extended abstract we present progress of RETROGRAM, a linguistic heritage project that focuses on the digitization and interpretation of pre-Illyrian Croatian grammar books with the aim to serve as a repository of digital Croatian grammars as well as to offer a model and develop processes on digitization of such works. Analyzing eight grammar texts published from the 17th until the 19th century, we developed two models: 1) a model for annotation of historical grammar and linguistic terminology, 2) a model for annotation of morphological paradigms. We composed a taxonomy consisting of 87 terms to be used in both models. To implement the models, we consulted the *TEI Guidelines*, the *de facto* standard in the digital humanities. Our first model adheres to the guidelines. However, our second model is a TEI-like model that we developed based on the dictionary module of the same guidelines. We hope that the morphological paradigm model will serve as a basis for the development of a TEI module for grammars, a model that is presently missing, but could be incorporated in the TEI infrastructure by expanding the dictionary module.

5. Acknowledgements

RETROGRAM is generously co-financed by the Croatian Science Foundation under the program "Research Projects" with grant agreement IP-2018-01-3585 and by the Institute of Croatian Language and Linguistics. We wish to thank all our research associates as well as Toma Tasovac for their feedback and help.

6. References

Francesco Maria Appendini and Ivana Lovrić Jović. 2022. *Appendinijeva Gramatika ilirskoga jezika: Jezična studija s prijevodom i transkripcijom uz faksimil*. Institut za hrvatski jezik i jezikoslovlje, Nacionalna i sveučilišna knjižnica u Zagrebu, Zagreb.

Ardelio Della Bella, Nives Sironić-Bonefačić, and Darija Gabrić-Bagarić. 2006. *Istruzioni grammaticali della lingua illirica, 1728: Gramatičke pouke o ilirskome jeziku*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

TEI Consortium (ed.). 2021a. 9 Dictionaries. In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 4.3.0. TEI Consortium. https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html.

TEI Consortium (eds.). 2021b. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.3.0. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Marijana Horvat. 2020. Istraživanje povijesti hrvatskoga jezika u digitalno doba. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(2):635–643.

Marijana Horvat and Martina Kramarić. 2021. Retro-Digitization of Croatian Pre-Standard Grammars. *Athens Journal of Philology,* 8(4):297–310.

Marijana Horvat and Ermina Ramadanović. 2012. *Jezikoslovni priručnik Blaža Tadijanovića Svašta po malo iliti kratko složenje imena, riči u ilirski i njemačkijezik (1761.)*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Bartol Kašić. 2002. *Institutiones linguae Illyricae/Osnove ilirskoga jezika*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Jakov Mikalja, Marijana Horvat,and Darija Gabrić-Bagarić. 2008. *Gramatika talijanska ukratko ili kratak nauk za naučiti latinski jezik*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Sanja Perić Gavrančić. 2020. *Latinska gramatika i hrvatski jezik Marijana Lanosovića: Povijesnojezična studija i transkripcija izvornika.* Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Ignacije Szentmártony. 2014. *Uvod u nauk o horvatskome jeziku*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Josip Voltić. 2016. *Grammatica Illirica/Ilirska gramatika. Reprint of the first edition (1803).* Institut za hrvatski jezik i jezikoslovlje, Zagreb.