Proceedings of the International Conference IMCS-60, MSU
October 10–13, 2024, Chisinau, Republic of Moldova

Digitization of Moldovan folklore texts using the HeDy platform

Vlada Colesnicova, Olesea Caftanatov, Svetlana Cojocaru, Alexandru Colesnicov, Ludmila Malahov

Abstract

This paper discusses the use of the digitization platform HeDy to produce electronic resources of Moldavian proverbs and sayings. It will promote the development and enrichment of annotated corpora and other electronic resources of philological data on regional folklore and nonstandard texts from Moldova. The results should be uniform in terms of formats and standards, in particular, to be incorporated into the Universal Dependency (UD) repository.

Keywords: Moldovan folklore, proverbs, sayings, OCR, Cyrillic script

1 Introduction: Digital Moldova

The Digital Transformation Strategy of the Republic of Moldova for 2023-2030 is designed to position the country as a regional leader in digital innovation and service delivery. By focusing on infrastructure, skills, and inclusive growth, Moldova aims to create a resilient digital economy that benefits all segments of society, ultimately enhancing the quality of life for its citizens.

It envisages that the Republic of Moldova "... will become an innovative and inclusive digital society with digital skills, with a developed digital infrastructure...". Artificial intelligence is considered one of the most important tools, which also contributes to natural language

©2024 by Vlada Colesnicova et al.

processing with applications for the Romanian language. The HeDy platform [1] supports the processing of printed or written documents from different periods of the Romanian language: digitization, transliteration, storage, analysis and subsequent use. The documents become accessible to a wide range of modern users, thus contributing to the enhancement of the value of our cultural heritage.

2 Key Features of the HeDy Platform

Features of the HeDy platfform are:

1. Digitization: advanced scanning technologies to convert physical documents into digital formats, ensuring clarity and detail. Used preservation techniques implement best practices for preserving the original documents while making digital copies.

2. Transliteration: employs sophisticated algorithms to transliterate texts accurately, allowing for easier reading and interpretation of historical documents, and facilitates the conversion of documents in various dialects and scripts, making them accessible to diverse audiences.

3. Storage.

3 Object of research: proverbs and sayings

Folklore is oral verbal creativity (myths, legends, fairy tales, etc.) and musical folk art, also verbal-musical and musical-choreographic folk works. More broadly, this includes all folk art, manifestations of the spiritual (and sometimes material) culture of the people: language, beliefs, rituals, crafts. The subject of scientific research since 1846, but registration and collecting occur even earlier.

We explore Moldovan proverbs and sayings published in the 60s-80s of the 20th century in the Cyrillic script, for example, [2]. These short instructive and entertaining statements have source in oral folklore, as

opposed to, for example, aphorisms. Proverbs are independent statements with a complete meaning. Sayings have meaning in a specific context. They have been found in written sources since the 16th-17th centuries.

The collections of proverbs and sayings from the Republic of Moldova represent a significant aspect of the nation's cultural heritage. Through their use of Cyrillic script and local subdialects, these collections not only preserve the linguistic diversity of the region but also reflect the values and experiences of the Moldavian people. As vital components of oral tradition, proverbs and sayings continue to resonate within contemporary culture, reminding us of the enduring power of folk wisdom.

Specific features of the mentioned sources include the Cyrillic script, and local (Moldovan) subdialects.

4 Technology of work

The processing includes the following stages: recognition, transliteration, verification, and creation of electronic resources. The first three stages are carried out on the HeDy platform.

5 Creation of electronic resources

Posting on the Internet is provided at all stages of work. For the correct introduction of these materials into scientific circulation and subsequent work of philologists, it is necessary to prepare data in certain formats. First of all, tokenization is performed: tokens are considered not only words, but also punctuation marks. Each token is assigned its characteristics (tags). These tokens and their characteristics are recorded in standard formats.

Morphological tagging focuses on the internal structure of words and their grammatical properties. This type of markup provides detailed information about each token, including part of speech, inflec-

tional features (tense, mood, number, gender, case), lemmas (the base form of words), affixes. etc.

Syntactic tagging addresses the relationships between words and their roles within sentences. This markup provides insights into the structure of phrases and clauses.

Acknowledgments. This work was supported by project 011301 "Information Systems based on Artificial Intelligence".

References

[1] T. Bumbu, L. Burtseva, S. Cojocaru, A. Colesnicov, and L. Malahov. *Digitization of Romanian Historical Printings.* Vladimir Andrunachievici Institute of Mathematics and Computer Science, Moldova State University, Chișinău, 2023 (Valinex), 176 p. ISBN 978-9975-68-497-2.

[2] *Proverbe și zicători.* Alcătuirea de E. Junghietu. Chișinău, "Știința", 1981. (Cyrillic script.)

Vlada Colesnicova[1,2], Olesea Caftanatov[1,3], Svetlana Cojocaru[1,4], Alexandru Colesnicov[1,5], Ludmila Malahov[1,6],

[1] Moldova State University,"V. Andrunachievici" Institute of Mathematics and Computer Science, Chisinau, Republic of Moldova
[2]ORCID: https://orcid.org/0009-0002-6801-193X
E–mail: vlada.colesnicova@math.usm.md
[3]ORCID: https://orcid.org/0000-0003-1482-9701
E–mail: olesea.caftanatov@math.usm.md
[4]ORCID: https://orcid.org/0009-0003-1025-5306
E–mail: svetlana.cojocaru@math.md
[5]ORCID: https://orcid.org/0000-0002-4383-3753
E–mail: acolesnicov@gmx.com
[6]ORCID: https://orcid.org/0000-0001-9846-0299
E–mail: ludmila.malahov@math.md