ROMAI J., v.14, no.2(2018), 33–40

REVITALIZATION OF THE RM FOLKLORIC TEXTS FROM THE SECOND HALF OF THE 20^{TH} CENTURY AND THEIR DIACHRONIC ANALYSIS

Tudor Bumbu, Olesea Caftanatov, Ludmila Malahov
Institute of Mathematics and Computer Science, Chişinău, Republic of Moldova
bumbutudor10@gmail.com, olesea.caftanatov@math.md, lmalahov@gmail.com

Abstract The aim of our work is revitalizing the folkloric texts of Republic of Moldova from the second half of the 20th century, furthermore actualizing the folkloric texts that are in Cyrillic writing for their future use in education. In addition, we intend to make a diachronic analysis between two periods (1960−1995 and 1996−2018).

Keywords: folkloric texts, Republic of Moldova, 20th century, Romanian Cyrillic writing, revitalization, diachronic analysis.
2010 MSC: 68T50.

1. INTRODUCTION

We assume that folklore reflects a certain vision of the populations life, beliefs and their feelings that can reach us through history. Our folk art that is manifested through songs, poetry, fairy tales, legends, proverbs and sayings, customs and traditions presents an invaluable wealth of treasure for all people who really love their homeland. Along the history, each nation created something that can be proud of, our country has something as well. Republic of Moldova is an area rich in authentic cultural traditions, unfortunately almost unknown beyond of Romanians borders. Furthermore, because of new digital trends, our native folk culture is being little by little forgotten. *Hence, we intend to revitalize the folkloric text to preserve our cultural heritage.* In Sec. 2, we present the importance of folkloric culture in our education and a brief comparison between three contemporary cultural environments. In Sec. 3, we present the digitization process of folkloric texts. In Sec. 4 of this paper, we make diachronic analysis between folklore from two different periods, and in the last Sec. 5, we share our ideas of creating an educational folkloric book based on our research.

2. THE ROLES OF THE NATIVE FOLK CULTURE

According to the Romanian Academician Solomon Marcus [1], pupils do not need complicated mathematics from the beginning of the school cycle, but to understand the environment, to discover the identity of their grandparents, to enrich their symbolic imagination. He claims that the folklore proposes models of morality and social behavior, much more appropriate that those offered by media and mass culture. Because the folk narratives do not describe a perfect world, impossible, broken from reality, as they would think, but it deploys to the pupils the whole range of human feelings, experiences and knowledge. For better understanding the importance of revitalization of the folkloric culture, we briefly presented a short comparison between three contemporary cultural environments. In Tab. 1, we summarize the few main characteristics of each cultural environments.

Table 1 Comparison between three contemporary cultural environments.

Pop culture	**Elite culture**	**Folk culture**
It is passed down by media.	It is passed down by media.	It is passed down verbally, rarely by media. Usually it is told among friends. And it is passed by visual perception.
It has short life span.	It has long life span.	It lives basically forever.
Its authors are traceable.	Its authors are known.	Its authors are unknown.

For our purpose, we used as base resource the book "Folclor din prile Codrilor" [2]. The processed ballads have been collected from rural population by Academy of Science folklorists during the 1965 till 1970 period and published in 1973 as an anthology.

3. THE DIGITIZATION OF NATIVE FOLKLORIC TEXT

Given the fact that, our main resources are books, we needed an OCR tool to convert image text into editable text. Thus, Optical Character Recognition performed by using ABBYY Finereader Professional 12 (FR). It is important to mention that in that period the Romanian language was using Cyrillic Alphabet. Because, this alphabet isnt integrate in FR, we created templates and added word dictionaries. In addition, we trained around 100 templates,

since; the most of letters can be find in Russian language with embedded templates in FR. The dictionary that we added has about 5,000 words, many of them are from other Cyrillic scripts that we recognized previously.

The digitization process included three steps.

Step 1. Scanning of the book [2].

Step 2. Recognition that was performed by using ABBYY FineReader 12 Professional software. Recognition accuracy is over 97% in words; 3% of erroneous words were corrected manually. Those errors were mainly because of low paper or print quality. As an example, the word in the left side of Fig.1 was recognized as the sequence in the right side of the same figure (see below).

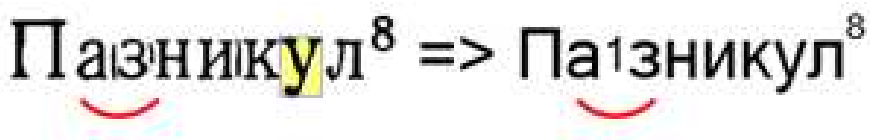

Fig. 1. An example of recognition error.

Step 3. Transliteration. Conversion from the Cyrillic alphabet into the Latin alphabet was performed with AAConv transliteration utility [3]. The transliteration accuracy was about 98%. To re-publish, text styles such as different fonts, italic, bold, etc. should be applied that was made manually. An example of these three steps is presented on Fig. 2.

Fig. 2. An example of the digitization process.

4. LEXICOGRAPHICAL DIACHRONIC ANALYSIS

Diachronic (across time) linguistics also known as historical linguistics investigates and describes the way in which languages change or maintain their structure over time. Diachronic analysis learns in which way language changes spread across spatial and temporal dimensions. It becomes evident that a language changes during the course of time when documents written in the same language but at different periods in time are subjected to examination. Orthographic and stylistic conventions which are characteristic of languages in their written form are the first changes to be observed over time. So our focalisation is on lexicographical diachronic analysis aspect of the language.

Diachronic analysis tools is one of the most popular today software of NLP. Their application supposes digitization of huge collections of historic texts and creation the corresponding corpora.

The presented paper focuses on resource specific features, which have to be implemented to maintain diachronic analysis. Actually any historical text based corpus can be made diachronic by slicing into time slices by natural time component.

We selected approx. 15,000 words from folkloric book [2] and 7,000 words from native lyrical songs to experiment with diachronic analysis. The folkloric book was printed using Moldavian Cyrillic Script, so we had to recognize and transliterate it into Latin Script with the above described tools. The next step was to verify the orthography and to correct the mistaken characters. This process was manual and it took a long time. In the next step we got rid of *stop words* by using a dedicated java program with a list of stop words manually selected form text. Some examples from the stop word list are: *"sunt"* ((are (III, plural and I, singular)) very common verb), *"de, în, pe, de pe"* (prepositions), *"şi, ca, să, ci, dar, de, fie, dacă, ori"* (conjunctions), etc. After this step we obtained a clean text which was involved in the next processes (see Fig. 3).

In the next step we tried to extract the words in common and for this purpose we used *latent Dirichlet allocation (LDA)* which is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. We used a version of this model implemented in the Python framework, namely GraphLab [4]. In the configuration of the model, we set only to select words form 1 topic and to iterate 200 times (optimal iterations for our data, we experimented also with 300, 400 and 500 iteration with bad results). In Fig. 4, we can see the learning results, and in Fig. 5 some of the resulted topic words.

```
Out[9]: dtype: dict
Rows: 2
[{'neamurile': 1L, 'nimica': 1L, 'budăi': 1L, 'ghietu': 1L,
'mărțișor': 1L, 'lăcata': 1L, 'dalici': 1L, 'uite': 6L, 'con
'temnița': 1L, 'mărire': 1L, 'dunăre': 1L, 'doctorii': 1L, '
1L, 'voinic': 3L, 'puiule': 1L, 'ploaia': 3L, 'capul': 6L, '
L, 'codrul': 3L, 'poată': 2L, 'cioc': 2L, 'căscată': 1L, 'în
1L, 'rămân': 2L, 'sărmana': 1L, 'șadă': 1L, 'partea': 1L, 'ș
'direcția': 1L, 'bunica': 1L, 'strigături': 4L, 'acușica': 1
'viței': 1L, 'găsi': 1L, 'minte': 1L, 'petru': 1L, 'viscoleș
l': 1L, 'tematică': 1L, 'petre': 2L, 'fugi': 2L, 'iubiții':
s': 1L, 'răcoare': 1L, 'tăcea': 1L, 'acasă': 10L, 'zbori': 1
ngheni': 37L, 'jurai': 1L, 'dispicătură': 1L, 'zbura': 5L, '
rângă': 1L, 'scăpa': 1L, 'dra': 1L, 'gogea': 1L, 'munteanu':
ludă': 1L, 'luminoasă': 2L, 'căutat': 2L, 'dovedim': 1L, 'li
L, 'împreună': 2L, 'nebun': 2L, 'îs': 6L, 'nănășelu': 1L, 'd
nschii': 1L, 'lungi': 2L, 'zidiau': 1L, 'me': 2L, 'aluatul':
l': 3L, 'străin': 3L, 'mp': 1L, 'zbera': 1L, 'împușcat': 1L,
oara': 1L, 'leacuri': 1L, 'răsplata': 1L, 'lungis': 1L, 'bun
```

Fig. 3. The array with the cleaned text.

```
In [10]: opere_topic_model = gl.topic_model.create(opere_text, num_topics=1, num_iterations=200)
Learning a topic model

       Number of documents         2

           Vocabulary size      3983

   Running collapsed Gibbs sampling

+-----------+---------------+----------------+-----------------+
| Iteration | Elapsed Time  | Tokens/Second  | Est. Perplexity |
+-----------+---------------+----------------+-----------------+
| 10        | 44.002ms      | 4470000        | 0               |
| 20        | 58.003ms      | 4470000        | 0               |
| 30        | 75.004ms      | 4470000        | 0               |
| 40        | 109.006ms     | 2235000        | 0               |
| 50        | 126.007ms     | 4470000        | 0               |
| 60        | 162.009ms     | 2235000        | 0               |
| 70        | 181.01ms      | 4470000        | 0               |
```

Fig. 4. LDA result.

In the extracted topic words are infiltrated some stop words which were not foreseen. For example the preposition *"într"*, used in one of this forms: *"într-o, într-un, într-adevăr"*. The final step was to measure the difference between terms form two different centuries and for this objective we used Levenshtein distance which is a string metric for measuring the difference between two sequences. Also, for the optimal approximation of the distance, Munkres algorithm [5] was implemented in Python. Some of the results are shown in Fig. 6.

```
a = [x['words'] for x in opere_topic_model.get_topics(output_type='topic_words',
for i in a:
    strn = '\n'.join(sorted(i, reverse=True))
    print strn
```

```
știi
șarpele
șa
întreabă
într
înscris
zunea
ziua
văzut
văd
vântul
vreau
vrea
vine
viața
verde
venit
```

Fig. 5. Some of the extracted topic words.

```
In [22]: def match_lists(l1, l2):

             matrix = [[levenshtein(i1, i2) for i2 in l2] for i1 in l1]
             indexes = Munkres().compute(matrix)
             for row, col in indexes:
                 yield l1[row], l2[col]

         for i1, i2 in match_lists(text1, text2):
         #     for z1 in str1:
         #         if i1 == z1 or i2 == z1:
         #             if levenshtein(i1,i2) < 3:
                       if i1 != i2:
                           print i1, '=>', i2
```

```
budăi => odăi
ghietu => bietu
dalici => mamiii
voinic => tainic
voit => vrut
codrul => coșul
```

Fig. 6. Some results after measuring the Levenstein distance between terms.

We can observe that the obtained results are not so good, but about a third from them (13/30) were what we needed. Below are listed the most impressive results:

- ghietu → bietu
- voit → vrut
- iaca → iată
- sara → seara
- sama → seama

- jele → jale
- caţi → cauţi
- mez → miez

While validating transliterated text we observed a lot of regional dialect words, that usually are used in countryside. Few of them we present below:

- a chicat → a căzut
- să şi fost → să fi fost
- Dumnezău → Dumnezeu
- parale → bani
- di şi → de ce
- jis → vis
- chirostrei → pirostrie

The result is a subject for philological expertise.

5. REVITALIZATION FOLKLORE FOR CHILDRENS BOOK

Like any other disciplines, folklore are very important in education the younger generation, because it perfectly fulfills the role of teaching the importance to pass dawn a culture morals, values, ancient beliefs, customs and rituals from one generation to the next one.

We suggest using our developed tools and getting the necessary resources in creating a folkloric book for schoolchildren. Thereby, we revitalize the folkloric text in order to preserve our cultural heritage. Additionally, we intend to use colorful illustrations for engaging kids to read the folkloric book.

Colorful illustrations are a main attraction in the childrens book because it motivates children to read the textual message. Through the images, the children learn and understand the world around them. For our folkloric book, we choose few text for each types of folklore such us: riddles, fairy tales, proverbs, folk tales, legends, lyrical songs, poetry, superstitions, urban legends, etc.

We organized a team of volunteer illustrators. Furthermore, each type of folklore will have their special style. We intend to combine nowadays art trends with folkloric text. We present below a page from book regarding riddles. Example is on Fig. 7.

Fig. 7. Illustrated riddles (ill. Alexandr Grosu).

In such way, we illustrated the other aspects of folklore like fable, legend, poetry, lyrical songs etc. Furthermore, we organized a workshop with schoolchildren dedicated native folklore, where we discussed the illustrations for the book. In addition, kids offered their drawings for some pages.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the process of digitization of native folkloric text with a recognition accuracy of 97% and transliteration accuracy of 98%. We intend to use the obtained resource to re-publish the book [2].

For this purpose, we involved a group of voluntaries to illustrate the folkloric texts. We also focused on lexicographical diachronic aspect of the language. The obtained result can be a subject for philological expertise and future research. We believe that nowadays it's important to revitalize and digitize the folkloric texts to preserve our cultural heritage.

References

[1] S. Marcus, *Semiotica folclorului. Abordare lingvistico-matematică*, Ed. Academiei, Bucureşti, 1975.

[2] G. G. Botezatu, H.M. Băeu, E.V. Junghientu, M.G. Savina, E.V. Tolstenco, A.S. Hîncu, V.A. Cirimpei i I.D. Ciobanu, *Folclor din prile Codrilor,* Academia de Ştiinţe a RSS Moldoveneşti, 1967.

[3] S. Cojocaru, A. Colesnicov, L. Malahov, *Digitization of Old Romanian Texts Printed in the Cyrillic Script,* Second International Conference on Digital Access to Textual Cultural Heritage DATeCH-2017, Goettingen, June 1-2, 2017, 143–148.

[4] Joseph Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, Carlos Guestrin, *PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs,* Proceedings of Operating Systems Design and Implementation (OSDI), 2012.

[5] Harold W. Kuhn, *The Hungarian Method for the assignment problem,* Naval Research Logistics Quarterly, 2: 8397, 1955.