Revitalizing the folkloric text of RM from the second half of the 20th century and their diachronic analysis

Bumbu Tudor, Caftanatov Olesea, Malahov Ludmila

Institute of Mathematics and Computer Science of ASM, Chişinău, Republic of Moldova
e-mail: bumbutudor10@gmail.com, olesea.caftanatov@math.md, lmalahov@gmail.com

Summary. 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).

We assume that folklore reflects a certain vision of the population's 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. For our purpose, we used as base resource the book *Folclor din părţile Codrilor* [1]. 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 U Finereader Professional 12 (FR). It is important to mention that in that period the Romanian language was using Cyrillic Alphabet. Because, this alphabet isn't 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 only exception is the letter that produce the sound ”gi”. The dictionary that we added has about 5000 words, many of them are from other Cyrillic scripts that we recognized previously.

Regarding, recognition accuracy is over 97% words the remaining 3% errors we corrected them manually. In order to convert from Cyrillic alphabet into Latin alphabet, we also used the AAConv Tool. This tool was developed within Institute of Mathematics and Computer Science. The transliteration accuracy is about 99%. For reediting text style such as, font, bolt italic, capital letters etc. we needed a bit of manual work. The obtained resource we intend to make a book. For its illustration were involved few volunteers. Finally, by using Machine-learning techniques such as, LDA latent dirichlet allocation and Euclidean distance, we analyzed the diachronic aspect of folklore text.

Bibliography

[1] G. G. Botezatu, H.M. Băeşu, E.V. Junghientu, M.G. Savina, E.V. Tolstenco, A.S. Hîncu, V.A. Cirimpei şi I.D. Ciobanu, *Folclor din părţile Codrilor*, Academia de Ştiinţe a RSS Moldoveneşti. 1967. p.356