WIIS2021

Proceedings of Workshop on Intelligent Information Systems

Vladimir Andrunachievici Institute of Mathematics and Computer Science
October 14-15, 2021, Chisinau

CZU 004+519.7(082)
P 93

VLADIMIR ANDRUNACHIEVICI INSTITUTE OF MATHEMATICS AND COMPUTER SCIENCE
5, Academiei str., Chisinau, Republic of Moldova, MD 2028
Tel: (373 22) 72-59-82, Fax: (373 22) 73-80-27
E-mail: imam@math.md
WEB address: http://www.math.md

Editors: Dr. I. Titchiev, Prof. C. Gaindric

Descrierea CIP a Camerei Naţionale a Cărţii

Proceedings of Workshop on Intelligent Information Systems WIIS2021, October 14-15, 2021, Chisinau / editors: I. Titchiev, C. Gaindric. – Chișinău : Vladimir Andrunachievici Institute of Mathematics and Computer Science, 2021 (Valinex SRL). – 214 p. : fig., fig. color, tab.

Antetit.: Vladimir Andrunachievici Inst. of Mathematics and Computer Science. – Referinţe bibliogr. la sfârşitul art. – 150 ex.

ISBN 978-9975-68-438-5.

004+519.7(082)

ISBN 978-9975-68-438-5

This issue is supported by the research project 20.80009.5007.22, "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

Proceedings of Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

On Symbolic Models Based on Markov Chains

Invited paper

Volodymyr G. Skobelev

Abstract

Symbolic models based on both discrete time and continuous time finite Markov chains, intended for studying behaviors of degrading systems, are presented from the single perspective. These models are defined according to the same schema. They provide an opportunity to determine three types of degrading systems, namely completely recoverable, partially recoverable and non-recoverable ones, and two types of critical sets of states. The problem of reachability the target set of states is analyzed.

Keywords: degrading systems, finite time horizon, finite Markov chains, discrete time, continuous time.

1 Introduction

Currently, finite stochastic models based on Markov Processes [1] are widely used for the analysis and behavior simulation of real systems. These models, in terms of frequencies, can be applied for investigating the critical sets of states reachability, implementing these or the others behaviors, and analyzing the degradation process of the studied system as a whole.

In doing so, the models based on Finite Markov Chains (FMC) are often used because of their simplicity and ease of handling. In this case, FMCs are used both with Discrete Time (DT) and with Continuous Time (CT). It should be noted that DT FMC models are used when the parameters measurements are carried out after the expiration of fixed

time intervals and the probabilities of state transitions are constant, while CT FMC models are used when state transition probabilities change over time.

It is evident that DT FMC models can also be used when state transition probabilities change over time. Indeed, it is sufficient to divide the time horizon into disjoint intervals, and on each of them apply the corresponding DT FMC model. This approach is equivalent to a piece-wise constant approximation of a continuous process.

Despite the numerous applications of FMC models, they are usually built for a specific problem with fixed numerical values of parameters. Therefore, the elaboration of symbolic FMC models (i.e. parameters are symbols) is relevant, since they give the possibility to apply analytical methods for studying behaviors of the classes of real systems with the same structure of transitions. These problems with respect to the analysis of the behavior of a degrading system (DS) have been solved in [2]-[4]. The given paper represents models elaborated in [3]-[4] from the single perspective.

2 Preliminary information

We will deal with FMC with the set of states $S_n = \{s_1, \ldots, s_n\}$ $(n \geq 2)$. We denote the DT FMC by $\mathcal{D}_n$ and the CT FMC by $\mathcal{C}_n$. These FMC can be determined and analyzed as follows:

1. The DT FMC $\mathcal{D}_n$ can be determined by the stochastic matrix

$$P_{\mathcal{D}_n} = \begin{array}{c|cccc} & s_1 & s_2 & \ldots & s_n \\ \hline s_1 & p_{11} & p_{12} & \ldots & p_{1n} \\ s_2 & p_{21} & p_{22} & \ldots & p_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_n & p_{n1} & p_{n2} & \ldots & p_{nn} \end{array},$$

i.e. $p_{ij} \geq 0$ $(i, j \in \mathbb{N}_n)$ and $\sum\limits_{j=1}^{n} p_{ij} = 1$ for all $i \in \mathbb{N}_n$. We note that p_{ij} $(i, j \in \mathbb{N}_n)$ is the probability of transition from the state s_i to the state

s_j in one step.

For any initial distribution of the states probabilities

$$\mathbf{p}_0 = (p_1^{(0)}, \ldots, p_n^{(0)}) \ \ (p_i^{(0)} \geq 0 \ \ (i \in \mathbb{N}_n), \ \ \sum_{i=1}^{n} p_i^{(0)} = 1)$$

the vector

$$\mathbf{p}_m = (p_1^{(m)}, \ldots, p_n^{(m)}) = \mathbf{p}_0 P_{\mathcal{D}_n}^m \ \ (m \in \mathbb{N}) \tag{1}$$

is the distribution of the states probabilities after m steps.

2. The CT FMC $\mathcal{C}_n$ can be determined by the transition rate matrix

$$\Lambda_{\mathcal{C}_n} = \begin{array}{c|cccc} & s_1 & s_2 & \ldots & s_n \\ \hline s_1 & -\sum\limits_{j \in \mathbb{N}_n \setminus \{1\}} \lambda_{1j} & \lambda_{12} & \ldots & \lambda_{1n} \\ s_2 & \lambda_{21} & -\sum\limits_{j \in \mathbb{N}_n \setminus \{2\}} \lambda_{2j} & \ldots & \lambda_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_n & \lambda_{n1} & \lambda_{n2} & \ldots & -\sum\limits_{j \in \mathbb{N}_n \setminus \{n\}} \lambda_{nj} \end{array},$$

where $\lambda_{ij} \geq 0$ $(i, j \in \mathbb{N}_n, \ i \neq j)$ is the rate of the departing from the state s_i and arriving in the state s_j.

Let $p_i(t)$ $(i \in \mathbb{N}_n)$ be the probability that the CT FMC $\mathcal{C}_n$ is in the state s_i at instant t. Then for any initial distribution of the states probabilities $\mathbf{p}_0$, the vector

$$\mathbf{p}(t) = (p_1(t), \ldots, p_n(t))$$

is the solution of the Chapman–Kolmogorov system of equations

$$\begin{cases} \frac{d\mathbf{p}(t)}{dt} = \mathbf{p}(t)\Lambda_{\mathcal{C}_n} \\ \mathbf{p}(0) = \mathbf{p}_0 \end{cases}, \tag{2}$$

i.e. a system of linear differential equations with constant coefficients with respect to the variables $p_i(t)$ $(i \in \mathbb{N}_n)$.

3 The symbolic FMC models for DS analysis

The symbolic FMC models for analyzing the behavior of a DS $\mathcal{S}_n$ with n stages of functionality have been developed in accordance with the same, in essence, scheme (the DT FMC model $\mathcal{D}_n$ in [3] and the CT FMC model $\mathcal{C}_n$ in [4]). This scheme can be presented as follows.

The states of the FMC $\mathcal{X}_n \in \{\mathcal{D}_n, \mathcal{C}_n\}$ represent the functionality stages of the DS $\mathcal{S}_n$ as follows. Let k $(1 < k \leq n)$ levels for functionality stages of the DS $\mathcal{S}_n$, enumerated by the elements of the set $\mathbb{N}_k$ in the decreasing order of the functionality be considered. So, the integer 1 corresponds to the completely functional stage, the integers $2, \ldots, k-1$ – to the partially functioning stages and the integer k – to the inoperable stage.

Let $f : S_n \to \mathbb{N}_k$ be the surjection such that if $i < j$ $(i, j \in \mathbb{N}_n)$, then $f(s_i) \leq f(s_j)$. We get the partition

$$\pi = S_n / \ker f = \{B_1, \ldots, B_k\},$$

where $B_1 = \{s_1\}$, $B_k = \{s_k\}$, and each block B_i $(i \in \mathbb{N}_{k-1} \backslash \{1\})$ consists of all states representing the stages with the same functionality level.

Let $x = p$ if $\mathcal{X}_n = \mathcal{D}_n$, and $x = \lambda$ if $\mathcal{X}_n = \mathcal{C}_n$. We set

$$S_n^{dsc}(s_r) = \{s_h \in \bigcup_{m=j+1}^{k} B_m \,|\, x_{rh} > 0\} \;\; (s_r \in B_j \;\; (j \in \mathbb{N}_{k-1})), \qquad (3)$$

and

$$S_n^{anc}(s_r) = \{s_h \in \bigcup_{m=1}^{j-1} B_m \,|\, x_{rh} > 0\} \;\; (s_r \in B_j \;\; (j \in \mathbb{N}_k \backslash \{1\})). \qquad (4)$$

It is supposed that the FMC $\mathcal{X}_n \in \{\mathcal{D}_n, \mathcal{C}_n\}$ satisfies to the following four conditions.

Condition 1. For all $j \in \mathbb{N}_{n-1}$ the equality $x_{nj} = 0$ holds.

Condition 2. For all $j \in \mathbb{N}_{k-1} \backslash \{1\}$ the equality $x_{rh} = 0$ holds for all states $s_r, s_h \in B_j$ $(r \neq h)$.

Condition 3. For all $j \in \mathbb{N}_{k-1}\backslash\{1\}$, if $x_{rn} = 0$ for all states $s_r \in B_j$, then $x_{hn} = 0$ for all states $s_h \in B_{j-1}$.

Condition 4. The equalities $S_n^{dsc}(s_r) \neq \emptyset$ $(s_r \in B_j \ \ (j \in \mathbb{N}_{k-1}))$ and $\bigcup\limits_{s_r \in B_j} (S_n^{dsc}(s_r) \cap B_{j+1}) = B_{j+1} \ \ (j \in \mathbb{N}_{k-1})$ hold.

Conditions 1 and 4 imply that the following two propositions are true.

Proposition 1. *The state s_n is the single absorbing state of the FMC $\mathcal{X}_n \in \{\mathcal{D}_n, \mathcal{C}_n\}$.*

Proposition 2. *In the FMC $\mathcal{X}_n \in \{\mathcal{D}_n, \mathcal{C}_n\}$ the state s_n is reachable from any state $s \in S_n\backslash\{s_n\}$.*

Formulae (3) and (4) and Conditions 1 – 4 imply the correctness of the following two definitions.

Definition 1. *In the FMC $\mathcal{X}_n \in \{\mathcal{D}_n, \mathcal{C}_n\}$ the critical set of states in the weak sense is*

$$S_n^{ws-cr} = \{s_h \in \bigcup_{m=1}^{j-1} B_m \,|\, s_n \in S_n^{dsc}(s_h)\}$$

and the critical set of states in the strong sense is

$$S_n^{ss-cr} = \{s_h \in \bigcup_{m=1}^{j-1} B_m \,|\, S_n^{dsc}(s_h) = \{s_n\}\}.$$

Definition 2. *Let the FMC $\mathcal{X}_n \in \{\mathcal{D}_n, \mathcal{C}_n\}$ be the model of the DS $\mathcal{S}_n$. Then:*

1. If $\mathcal{S}_n$ is a completely recoverable DS, then the disequalities $S_n^{anc}(s_r) \neq \emptyset$ $(s_r \in B_j)$ hold for all $j \in \mathbb{N}_{k-1}\backslash\{1\}$.

2. If $\mathcal{S}_n$ is a partially recoverable DS, then there exists some integer $l \in \mathbb{N}_{k-1}\backslash\{1\}$ such that the disequalities $S_n^{anc}(s_r) \neq \emptyset$ $(s_r \in B_j)$ hold for each integer $j \in \mathbb{N}_{k-1}\backslash\mathbb{N}_l$, and the equalities $S_n^{anc}(s_r) = \emptyset$ $(s_r \in B_j)$ hold for each integer $j \in \mathbb{N}_l\backslash\{1\}$.

3. If $\mathcal{S}_n$ is a non-recoverable DS, then the equalities $S_n^{anc}(s_r) \neq \emptyset$ $(s_r \in B_j)$ hold for all $j \in \mathbb{N}_{k-1}\backslash\{1\}$.

Examples of the symbolic FMC models for these three types of DS have been presented in [3] (the DT FMC models $\mathcal{D}_n$) and [4] (the CT FMC models $\mathcal{C}_n$).

4 Analysis of the FMC model $\mathcal{X}_n \in \{\mathcal{D}_n, \mathcal{C}_n\}$

It is supposed that for the symbolic FMC model $\mathcal{X}_n$ the initial probability distribution of the states is $\mathbf{p}_0 = (1, \underbrace{0, \ldots, 0}_{n-1 \ times})$, i.e. that initially the analyzed DS $\mathcal{S}_n$ is in the completely functional stage. Besides, the behavior of the investigated DS $\mathcal{S}_n$ is analyzed within the finite time horizon, namely, l state transitions for the DT FMC model $\mathcal{D}_n$ and interval $[0, T]$ for the CT FMC model $\mathcal{C}_n$. The target set of states S_n^{trgt} for the model $\mathcal{X}_n$ is defined as an element of the set $\{\{s_n\}, S_n^{ws-cr}, S_n^{ss-cr}\}$.

4.1 The DT FMC model $\mathcal{D}_n$

Using (1), for each $m \in \mathbb{N}_l$, parametric expressions for the probabilities of reachability the target set of states

$$\mathbb{P}_{s_1, S_n^{trgt}}(m) = \sum_{s_i \in S_n^{trgt}} p_i^{(m)} \ \ (m \in \mathbb{N}_l) \tag{5}$$

can be computed. We note that the parametric expressions (5) are the functions of the parameters p_{ij} $(i, j \in \mathbb{N}_n)$.

Therefore, partial derivatives

$$\frac{\partial \mathbb{P}_{s_1, S_n^{trgt}}(m)}{\partial p_{ij}} \ \ (m \in \mathbb{N}_l) \tag{6}$$

characterizing the probabilities $\mathbb{P}_{s_1, S_n^{trgt}}(m)$ $(m \in \mathbb{N}_l)$ variations relatively to each parameter p_{ij} $(i, j \in \mathbb{N}_n)$ variation can be computed.

Moreover, for each $m \in \mathbb{N}_l$, such that $\mathbb{P}_{s_1, S_n^{trgt}}(m) \neq 0$, the rates of change

$$\mathbb{P}^{-1}_{s_1, S_n^{trgt}}(m) \frac{\partial \mathbb{P}_{s_1, S_n^{trgt}}(m)}{\partial p_{ij}} \ \ (i, j \in \mathbb{N}_n) \tag{7}$$

for the probabilities $\mathbb{P}_{s_1,S_n^{trgt}}(m)$ $(m \in \mathbb{N}_l)$ relatively to each parameter p_{ij} $(i,j \in \mathbb{N}_n)$ variation can be computed.

The parametric expressions (6) and (7) form some base for statistical simulation [5] of the DS $\mathcal{S}_n$ behavior under parameters' variation.

For the DT FMC model $\mathcal{D}_n$, the reachability problem for the target set of states when the numeric values of the parameters p_{ij} are given has been solved in [3] on the base of bounded analysis [6]. This method is as follows.

Let $\mathcal{P}(s_1, S_n^{trgt}, \gamma)$ be the property:

$$\text{"the inequality } \sum_{i=1}^{l} \mathbb{P}_{s_1,S_n^{trgt}}(i) \leq \gamma \text{ holds"},$$

where γ $(0 < \gamma < 1)$ is the given threshold.

Checking the property $\mathcal{P}(s_1, S_n^{trgt}, \gamma)$ consists of computing sequentially the values $\sum_{i=1}^{m} \mathbb{P}_{s_1,S_n^{trgt}}(i)$ $(m = 1, \ldots, l)$ and checking for each computed value whether the inequality $\sum_{i=1}^{m} \mathbb{P}_{s_1,S_n^{trgt}}(i) \leq \gamma$ holds.

If all these inequalities hold, then the property $\mathcal{P}(s_1, S_n^{trgt}, \gamma)$ is true. Otherwise, let m $(m \in \mathbb{N}_l)$ be the smallest integer such that $\sum_{i=1}^{m} \mathbb{P}_{s_1,S_n^{trgt}}(i) > \gamma$. Then some minimal by cardinality subset $\mathcal{S}$ of the set S_n^{trgt} such that $\sum_{i=1}^{m} \mathbb{P}_{s_1,\mathcal{S}}(i) > \gamma$ is computed. The set of transitions from s_1 to $\mathcal{S}$ in no more than m steps is called a counterexample [7].

4.2 The CT FMC model $\mathcal{C}_n$

When we deal with a completely recoverable or a partially recoverable DS $\mathcal{S}_n$, the Chapman–Kolmogorov system of equations (2) is a parametric system of linear differential equations of a general form with constant coefficients. Its solution can be found by reducing it either to a parametric system of algebraic equations by using direct and inverse Laplace transformations, or to a parametric linear homogeneous differential equation of the n -th order with constant coefficients. When we deal with a non-recoverable DS $\mathcal{S}_n$, the Chapman–Kolmogorov system

of equations (2) is a triangular parametric system of linear differential equations with constant coefficients. Its solution can be found by sequentially solving the equations, starting with the first equation. This method has been illustrated in [4].

Let $\mathbf{p}(t) = (p_1(t), \ldots, p_n(t))$ be the solution of the system of equations (2). We note that each $p_i(t)$ $(t \in [0, T])$ is the function of time and the parameters λ_{ij} $(i, j \in \mathbb{N}_n)$. So, partial derivatives

$$\frac{\partial p_i(t)}{\partial t} \ (t \in [0, T]) \tag{8}$$

characterizing the probabilities $p_i(t)$ $(i \in \mathbb{N}_n)$ as functions of time via the relations between the parameters can be computed.

Also there can be computed parametric expressions for the probabilities of reachability the target set of states

$$\mathbb{P}_{s_1, S_n^{trgt}}(t) = \sum_{s_i \in S_n^{trgt}} p_i(t) \ (t \in [0, T]). \tag{9}$$

So, partial derivatives

$$\frac{\partial \mathbb{P}_{s_1, S_n^{trgt}}(t)}{\partial \lambda_{ij}} \ (t \in [0, T]) \tag{10}$$

characterizing the probabilities $\mathbb{P}_{s_1, S_n^{trgt}}(t)$ $(t \in [0, T])$ variations relatively to each parameter λ_{ij} $(i, j \in \mathbb{N}_n)$ variation can be computed.

Moreover, when $\mathbb{P}_{s_1, S_n^{trgt}}(t) \neq 0$, the rates of change

$$\mathbb{P}^{-1}_{s_1, S_n^{trgt}}(t) \frac{\partial \mathbb{P}_{s_1, S_n^{trgt}}(t)}{\partial \lambda_{ij}} \ (i, j \in \mathbb{N}_n) \tag{11}$$

for the probabilities $\mathbb{P}_{s_1, S_n^{trgt}}(t)$ $(t \in [0, T])$ relatively to each parameter λ_{ij} $(i, j \in \mathbb{N}_n)$ variation can be computed.

The parametric expressions (8)–(11) form some base for statistical simulation of the DS $\mathcal{S}_n$ behavior in time and under parameters' variation.

For the CT FMC model $\mathcal{C}_n$, the reachability problem for the target set of states when the numeric values of the parameters λ_{ij} are given has been solved in [4]. We note that substituting these values into the parametric expressions $p_i(t)$ $(i \in \mathbb{N}_n)$, we get $p_i(t)$ $(i \in \mathbb{N}_n)$ as functions of time only.

The solutions proposed in [4] are as follows.

Let $S_n^{trgt} = \{s_n\}$. Then $\mathbb{P}_{s_1,\{s_n\}}(t) = p_n(t)$ $(t \in [0,T])$ is a strictly monotone increasing function. So, the problem of analysis of the target set reachability can be formulated as follows.

Problem 1. *Let the numbers* ε $(0 < \varepsilon < 1)$ *and* τ $(0 < \tau < 0.5T)$ *be given. It is necessary to find* $t_0 \in [0,T]$ *such that*

$$\begin{cases} p_n(\max\{0, t_0 - \tau\}) < \varepsilon \\ p_n(\min\{T, t_0 + \tau\}) \geq \varepsilon \end{cases}.$$

The solution to this problem is as follows.

If $p_n(T) < \varepsilon$, then the required value $t_0 \in [0,T]$ does not exist. Otherwise, the bisection method can be applied to find the required value $t_0 \in [0,T]$ (see Algorithm 1 and Theorem 1 in [4]).

Let $S_n^{trgt} \in \{S_n^{ws-cr}, S_n^{ss-cr}\}$. There is no guarantee that $\mathbb{P}_{s_1,S_n^{trgt}}(t)$ $(t \in [0,T])$ is a strictly monotone function. So, the problem of analysis of the target set reachability can be formulated as follows.

Problem 2. *Let the number* ε $(0 < \varepsilon < 1)$ *and the positive integer* k *be given. It is necessary to find the set*

$$\mathcal{T}_{s_1,S_n^{trgt}}(k,\varepsilon) = \{h2^{-k}T | h \in \{0,1,\ldots,2^k\} \& \mathbb{P}_{s_1,S_n^{trgt}}(h2^{-k}T) \geq \varepsilon\}$$

in the explicit form.

The solution to this problem is as follows.

The values $\mathbb{P}_{s_1,S_n^{trgt}}(h2^{-k}T)$ $(h \in \{0,1,\ldots,2^k\})$ are computed sequentially and each of them is compared with the number ε (see Algorithm 2 and Theorem 2 in [4]).

5 Discussion

In the given paper the symbolic DT and CT FMC models, intended for DS behavior analysis within the finite time horizon, that have been elaborated in [3]-[4] are presented from the single perspective.

The presented symbolic FMC models give the possibility to obtain analytical expressions for the probabilities of being the analyzed DS in any functionality stage at any instant. In particular, analytical expressions for the probability of reachability the set of the target states at any instant can be obtained.

The above pointed analytical expressions can be used for investigation of the probabilities of being the analyzed DS at this or another functionality stage as time functions and as the parameter functions, both. This is especially important in the process of designing DSs.

Besides, the above pointed analytical expressions can be used for statistical simulation [5] of the analyzed DS behavior. In particular, through Monte Carlo simulation [8], these analytical expressions can also be used for searching the most acceptable vectors of numerical parameters values for the designed DS. In this case some relevant logical Model-Checking formalisms [9] can be used.

6 Conclusion

In the given paper the symbolic DT and CT FMC models intended for unified analysis of the behavior of the class of DSs $\mathcal{S}_n$ with the same transition structure within the finite time horizon that have been elaborated in [3]-[4] are presented from the single perspective.

The development of logical Model-Checking formalisms for the process of DS analysis is some direction for future research.

Another direction for future research is the development and implementation of statistical simulation methods for analysis of DS behavior on the base of the parameters variation.

References

[1] O.S. Ibe. *Markov processes for stochastic modelling. Second edition.* Elsevier Inc., 2013.

[2] V.G. Skobelev, V.V. Skobelev. *On degrading systems modelling via finite Markov Chains.* In: Proc. of the Workshop on Intelligent Information Systems, Chisinau, Moldova, 2020, pp. 189–195.

[3] V.G. Skobelev, V.V. Skobelev. *A general finite Markov chain model for degrading systems analysis.* IOSR Journal of Mathematics, vol. 17, issue 1, ser. III, 2021, pp. 19–29.

[4] V.G. Skobelev, V.V. Skobelev. *A symbolic continuous time Markov Chain model for degrading systems analysis.* IOSR Journal of Mathematics, vol. 17, issue 3, ser. III, 2021, pp. 62–72.

[5] D.R. Cox. *Role of models in statistical analysis.* Statistical Science, vol. 5, issue 2, 1990, pp. 169–174.

[6] A. Biere, A. Cimatti, E. M. Clarke, O. Strichman, and Y. Zhu, *Bounded model checking.* Advances in Computers, vol. 58, 2003, pp. 118–149.

[7] N.E. Abraham, B. Becker, C. Dehnert, et al. *Counterexample generation for discrete-time Markov models: an introductory survey.* Lecture Notes in Computer Science, vol. 8483, 2014, pp. 65–121.

[8] R.Y. Rubinstein, D.P. Kroese. *Simulation and the Monte Carlo Method, Third Edition.* John Wiley & Sons, Inc., 2017.

[9] A. Azis, K. Sanwal, V. Singhal, R.K. Brayton. *Model-Checking continuous-time Markov chains.* ACM Transactions on Computational Logic, vol. 1, issue 1, 2000, pp. 162–170.

Volodymyr G. Skobelev

V. M. Glushkov Institute of Cybernetics of NAS of Ukraine,
40 Glushkova ave., Kyiv, Ukraine, 03187
E–mail: skobelevvg@gmail.com

Natural Language Processing for Book Recommender Systems

Invited paper

Diana Inkpen

Abstract

Reading has benefits for individuals and societies, yet studies show that reading declines, especially among the young. Recommender systems can help stop this decline. There is a lot of research regarding literary books using natural language processing methods, but the analysis of textual book content to improve recommendations is relatively rare. We propose content-based recommender systems that extract elements learned from book texts to predict readers' future interests. Our first approach recommends books after learning their authors' writing style. The second approach uses lexical, syntactic, stylometric, and fiction-based features that might play a role in generating high-quality book recommendations. We evaluated both approaches according to a top-k recommendation scenario. They give better accuracy when compared with state-of-the-art content and collaborative filtering methods.

Our content-based systems suffer from the new user problem, well-known in the field of recommender systems, that hinders their ability to make accurate recommendations. Therefore, we propose an additional topic model component that addresses the issue by using the topics learned from a user's shared text on social media, to recognize their interests and map them to related books. This is also a novelty in the field of book recommender systems.

Diana Inkpen

University of Ottawa, Canada
E-mail: diana.inkpen@uottawa.ca

Proceedings of the Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

Efficient hardware implementations of membrane computing models

Invited paper

Sergey Verlan

Abstract

Unconventional computing models allow a different approach to the solution of numerous problems, especially in the field of optimization and parallel programming. In most cases, underlying models are massively parallel and cannot be implemented efficiently in software using current computer architectures.

Hence, it becomes interesting to consider specialized hardware (usually based on FPGA) in order to obtain a really parallel execution of corresponding models. To achieve efficient implementations, we analyzed the operations proposed by FPGA hardware and proposed restrictions to the theoretical model. This allowed obtaining efficient implementations featuring a speed-up of order 10^5 with respect to software implementations (and running $\sim 10^8$ steps per second).

A particularity of our approach is the use of results from the theory of formal languages to obtain the above speed-ups. We shall give examples (concerning the implementation of different variants of P systems), where the efficiency of the implementation is a consequence of the reduction of the simulation problem to the problem of the number of words in a regular language. Another example ties a high-performance implementation of numerical P systems with the ability to express the problem in Presburger arithmetic. We also show that in this last case the model corresponds to a system of non-linear difference equations and we apply this correspondence to the design of high-speed robot motion controllers.

Finally, we will discuss the link between our results and data flow synchronous programming languages, more particular LUSTRE, as well as questions related to the verification of the obtained models.

Sergey Verlan
University of Paris Est Creteil, France
E-mail: verlan@u-pec.fr

Cognitive Distributed Computing System Based on Temporal Logic

Victor Ababii, Viorica Sudacevschi, Silvia Munteanu, Victoria Alexei, Radu Melnic, Ana Turcan, Vadim Struna

Abstract

The paper presents the results of the conceptual and structural design of a cognitive system of distributed computing based on temporal logic. The cognitive system has a Multi-Agent structure that forms a mesh network with broadcast communication, which ensures the organization of knowledge exchange between them. Functional elements and temporal logical operators are defined in the form of mathematical models, which allows their implementation based on hardware devices or software products. The functionality of the temporal logic is determined by the time function that calculates the credibility coefficient of the event and its influence on the decisions taken by the Agents.

Keywords: cognitive system, distributed computing, Multi-Agent system, collective decision making, temporal logic, knowledge base.

1. Introduction

Conceptually, the cognitive system has a complex structure that has the ability to learn and develop new knowledge. A cognitive system can be presented in the form of a person, a group of people, an organization, an Agent, a computer, or a combination of these. The purpose of the cognitive system is to provide services to enhance the cognitive abilities of human agents (human intelligence) to select optimal solutions in solving problems in various fields of science. [1, 2, 3].

As mentioned in papers [4, 5, 6], the most effective solutions in solving complex problems are offered by distributed computing, parallel computing, and cloud computing systems. This work provides knowledge

in modeling distributed computing systems, cloud platforms, clustering technologies. Methods for improving performance in terms of scalability and reliability are analyzed. All the characteristics mentioned above are also specific to the design of Multi-Agent systems [7, 8] or of collective/collaborative decision-making systems [9], based on Artificial Intelligence.

The cognitive process is an evolutionary process that involves the analysis of a set of knowledge already known so far in order to generate new knowledge based on which optimal decisions can be made. This mechanism can very well be achieved by applying temporal logic [10, 11, 12], which allows highlighting the temporal relationship between past, present, and future.

2. Formulation of the design and research problem

Modern technologies offer a wide range of solutions that allow the development of distributed computing cognitive systems that ensure the solution of complex problems for different application areas. It can be mentioned the papers [13, 14, 17, 19], in which there were researched Multi-Agent systems oriented towards collective calculation, and [15, 16, 18], where cognitive systems based on knowledge with application in various fields are researched.

This paper proposes the conceptual and structural design of a cognitive system of distributed computing based on temporal logic. The architecture of the system presents lots of Agents that form a mesh network with broadcast communication, and it is oriented towards making collective decisions based on the knowledge accumulated over time. At the basis of the process of knowledge formation, there is the set of rules defined by temporal logic that establishes the connection between past, present, and future.

3. Synthesis of the Distributed Cognitive Calculation System

There is defined the distributed system of collective calculation consisting of the set of Agents $A = \left\{ A_i, i = \overline{1, I} \right\}$. The functional model of an Agent is defined by expression (1), its diagram is presented in figure 1:

$$A_i = \left\{ KB[T], LC : (Ev, Pp, TLP, DMP, Ac) \right\}_i, \quad (1)$$

where: $KB[T]$ – Knowledge Base at a time T; LC – Logic Control for synchronizing the data processing operations performed by the Agent; Ev – the set of external events perceived by the Agent and generated by the activity environment; Pp – Pre-processor for conditioning the signals generated by external events; TLP – Temporal Logic Processor that through the application of Temporary Operators processes the data from the knowledge base $KB[T]$ and input data generated by external events to generate new knowledge $KB[T+1]$; DMP – Decision-Making Processor that generates new decisions based on knowledge; Ac – Actions on the activity environment.

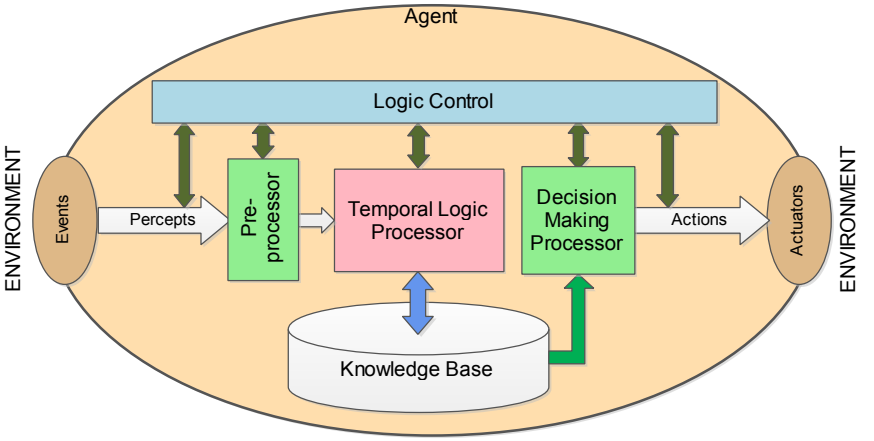

Figure 1. Agent Diagram.

Definition of functional elements of the system:

Event: $Ev[T]_i = \left\{ X[T], KB[T] \right\}_i, i = \overline{1, I}$, where $Ev[T]_i$ – the set of events that took place at the time $[T-1, T]_i$, but perceived by the Agent at the time T; $X[T]_i$ – the set of events generated by the activity

environment; $KB[T]_i$ – the multitude of knowledge received from all Agents involved in decision-making.

Decision:

$D[T]_i = \{DMP : KB[T] \to (Ac, Qr, Dec, Con)\}_i, i = \overline{1, I}$, where $D[T]_i$ – the set of decisions made by the Agent at the time T; Ac – decisions on action on the activity environment; Qr – generating questions to all the Agents; Dec – declaratory decisions communicated to all the Agents; Con – confirmatory decisions communicated to all the Agents.

Action: $Ac[T]_i = \left\{DMP : KB[T] \to \left(Ac[T]_{i,1}, ..., Ac[T]_{i,J_i}\right)\right\}, i = \overline{1, I}$,

where $Ac[T]_i$ is the set of actions generated by Agent A_i on the activity environment as a result of processing the decision-making block DMP of data from the knowledge base $KB[T]$.

Question: $Qr[T]_i = \left\{DMP : KB[T] \to \left(Qr[T]_{i,1}, ..., Qr[T]_{i,J_i}\right)\right\}, i = \overline{1, I}$,

where $Qr[T]_i$ is the multitude of questions addressed by the Agent A_i to the set of the Agents A as a result of processing the decision-making block DMP of data from the knowledge base $KB[T]$.

Declaration: $Dec[T]_i = \left\{DMP : KB[T] \to \left(Dec[T]_{i,1}, ..., Dec[T]_{i,J_i}\right)\right\}, i = \overline{1, I}$,

where $Dec[T]_i$ is the set of declarative information transmitted by the Agent A_i to the set of Agents A.

Confirmation:

$Con[T]_i = \left\{DMP : KB[T] \to \left(Con[T]_{i,1}, ..., Con[T]_{i,J_i}\right)\right\}, i = \overline{1, I}$, where $Con[T]_i$ is the set of confirmatory information transmitted by the Agent A_i to the group of Agents A as a result of processing questions $Qr[T]$ generated by the Agents A.

Temporal Logic Processor: $TLP_i = \left\{ O(\tau)_{i,1}, ..., O(\tau)_{i,J_i} \right\}, i = \overline{1,I}$,

where $O(\tau)_i$ is the set of temporal logic operators defined for the Agent A_i and implemented on a processor basis. The application of the set of temporal logic operators determines the cognitive properties of the distributed computing system.

Operator: $O(\tau)_i : \left\{ Ev[T]_i, KB[T]_i \right\} \to \left\{ KB[T+1]_i \right\}, i = \overline{1,I}$,

where: $Ev[T]_i$ is the set of events perceived from the activity environment; $KB[T]_i$ – the content of the knowledge base at the time T; and $KB[T+1]_i$ – knowledge base after application of the temporal logic operator $O(\tau)_i$.

The time function for temporal logic operators $O(\tau)_i$ is determined by the expression (2):

$$x(t) = x[T] / \left(k + t^2 / s \right), t = \overline{T, \infty}, \quad (2)$$

where: $x(t)$ is the value of the decisional influence (credibility) of the event on the content of the knowledge base $KB[T]$; k – the attenuation coefficient of credibility; s – stability coefficient of decision-making influence (decision-making credibility).

The structure and basic components of the operator are determined by the expression (3):

$$O(\tau) : \left\{ Op_1, Op_2, ..., Op_J \right\}, \quad (3)$$

where $Op_j, j = \overline{1,J}$ is the set of operands that are part of the operator structure $O(\tau)$.

The format of the operand is determined by the expression (4):

$$Op(\tau) = \left\{ Name, X[T], k, s \right\}, \quad (4)$$

where $Name$ is the name of the operand or its content (State, Question, Confirmation, Statement).

Table 1 presents the initial data for model validation (2). The modeling results are shown in Figure 2.

Table 1. Initial data for model validation (2)

Graphic number	$x[T]$	k	s
1	1	1	10
2	1	2	10
3	1	1	5
4	1	1	20
5	1	1	40

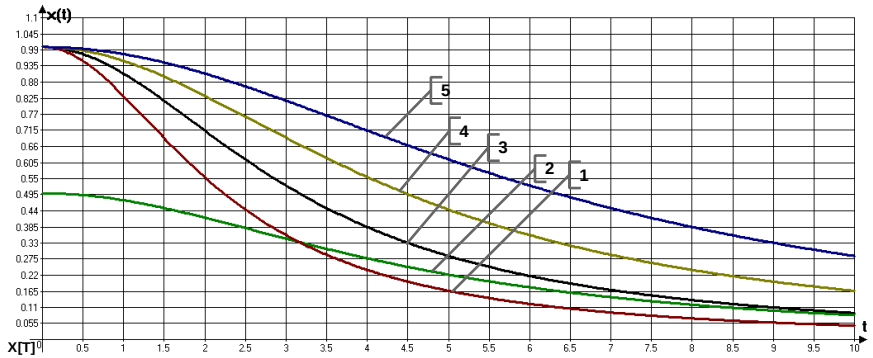

Figure 2. Model validation results (2).

4. Defining Temporal Logic Operators

The list of temporal logical operators is determined by the functionality and field of activity of the distributed cognitive computing system. In the following some examples of temporal logical operators are presented:

1) $O\left(\vee^{\tau}\right)=\max\left\{Op_1,Op_2,...,Op_J\right\}$;

2) $O\left(\wedge^{\tau}\right)=\min\left\{Op_1,Op_2,...,Op_J\right\}$;

3) $O\left(\neg^{\tau}\right)=NOT\left\{Op_1,Op_2,...,Op_J\right\}$;

4) $O\left(\cup^{\tau}\right)=\bigcup\left\{Op_1,Op_2,...,Op_J\right\}$;

5) $O\left(\cap^{\tau}\right)=\bigcap\left\{Op_1, Op_2, ..., Op_J\right\}$;

6) $O\left(\sum^{\tau}\right)=\sum\left\{Op_1, Op_2, ..., Op_J\right\}$;

7) $O\left(\prod^{\tau}\right)=\prod\left\{Op_1, Op_2, ..., Op_J\right\}$;

The number and complexity of operators for temporal logic can be extended in relation to the solved problem.

5. Conclusion

This paper presents the results of the conceptual and structural design of a distributed cognitive computing system based on temporal logic. The system is defined as a group of Agents that forms a network of computing devices with a mesh topology. The communication between the Agents ensures the organization of the knowledge exchange which allows the implementation of the calculation models with a collective decision.

The functionality of the Agents is based on the application of temporal logic models and includes: operations to perceive the state of the activity environment, communication with other Agents for the purpose of knowledge exchange, updating knowledge, calculating decisions and acting with these decisions on the activity environment, or their communication to other Agents.

The paper proposes the synthesis of the distributed system of cognitive calculation, which includes: the functional model of Agents and its diagram, the functional elements are defined in the form of mathematical models (Event, Decision, Action, Question, Declaration, Confirmation, and Temporal Logic Processor).

A temporal Logic Processor is defined as a set of operators that performs operations on a set of operands. The purpose of the operators is to update the knowledge base, thus offering cognitive capabilities for the system. An operand is a functional description of an event and includes its name or content, the initial state of the event, its credibility attenuation coefficient, and its credibility stability coefficient.

In order to validate the model for calculating the credibility coefficient of the event, its modeling was performed for different attenuation and stability coefficients. The results are presented in the form of graphs.

The implementation of the Agents in the form of hardware computing architectures and software products is planned for the future.

The results of this paper will be applied in the development of Multi-Agent systems with calculation and collective decisions.

Acknowledgments. The results presented in this paper are part of the research led at the Department of Computer Science and Systems Engineering. The functional tests of the developed models were performed with the technical and technological support offered by the Laboratory "Robotic and Mechatronic Systems".

References

[1] A. Gomila and V.C. Muller. *Challenges for Artificial Cognitive Systems*. Journal of Computer Science, No. 13, pp. 453-469, 2012.

[2] A. Langus and M. Nespot. *Cognitive Systems Struggling for Word Order*. Cognitive Psychology, No. 60, pp. 291-318, 2010. doi:10.1016/j.cogpsych.2010.01.004.

[3] L. Ogiela, R. Tadeusiewicz, and M.R. Ogiela. *Cognitive Techniques in Medical Information Systems.* Computers in Biology and Medicine, No. 38, pp. 501-507, 2008. doi:10.1016/j.compbiomed.2008.01.017.

[4] K. Hwang, G.C. Fox, and J.J. Dongarra. *Distributed and Cloud Computing*. ELSEVER, 670p., 2012, ISBN: 978-0-12-385880-1.

[5] B. Talekar, S. Chaudhari, and V. Shinde. *Distributed Computing Challenges.* IOSR Journal of Computer Engiinering, Vol. 16, Issue 2, pp. 28-31. 2014, ISSN: 2278-8727.

[6] K.N. Ahire. *Distributed Computing – Future and Applications.* International Research Journal of Engineering and Technology, Vol. 04, Issue 10, pp. 2005-2007. 2017, ISSN: 2395-0072.

[7] A. Byrski, R. Drezewski, L. Siwik, and M. Kisiel-Dorohinski. *Evolutionary Multi-Agent Systems.* The Knowledge Engineering Review, Vol. 30, Issue 2, pp. 171-186. 2015. DOI: 10.1017/S0269888914000289.

[8] J. Rocha. *Multi-Agent Systems.* Published by InTechOpen. 218p., 2017. ISBN 978-953-51-3535-7.

[9] L. Chen, C. Huepe, and T. Gross. *Adaptive Network Models of Collective Decision Making in Swarming Systems.* Phys. Rev. E 94, 022415 (2016). DOI: 10.1103/PhysRevE.94.022415.

[10] T. Agotnes, W. Hoek, J.A. Rodriguez-Aguilar, C. Sierra, and M. Wooldridge. *A Temporal Logic of Normative Systems.* Trens in Logic, No. 27, pp. 11-48, Springer, 2008.

[11] F. Laroussinie, N. Markey, and P. Schnoebelen. *Temporal Logic with Forgettable Past.* In: The Proceedings of the 17th Annual IEEE Symposium on Logic in Computer Sciences (LICS'02), pp. 383-392. 2002. ISBN: 0-7695-1483-9. DOI: 10.1109/LICS.2002.1029846.

[12] A. Rodionova, E. Bartocci, D. Nickovic, and R. Grosu. *Temporal Logic as Filtering.* In: The Proceedings of the 19th International Conference on Hybrid Systems: Computation and Control (HSCC'16), April 12-14, 2016, Vienna, Austria, pp. 11-20. DOI: 10.1145/2883817.2883839.

[13] V. Ababii, V. Sudacevschi, and Gh. Safonov. *Designing a Collective Agent for synthesis of Adaptive Decision-Making Systems.* Sciences of Europe (Praha, Chech Republic), Vol 1, No 17(17), 2017, pp. 70-75, ISSN: 3162-2364.

[14] V. Ababii, V. Sudacevschi, R. Melnic, and S. Munteanu. *Multi-Agent System for Distributed Decision-Making.* National Science Journal (Ekaterinburg, Russia), Vol 2, No 45, 2019, pp. 19-23, ISSN: 2413-5291. DOI: 10.31618/nas.2413-5291.2019.2.45.

[15] V. Ababii, V. Sudacevschi, R. Braniste, A. Turcan, C. Ababii, and S. Munteanu. *Adaptive computing system for distributed process control.* International Journal of Progressive Sciences and Technologies. Vol. 22, No 2, September 2020, pp. 258-264. ISSN: 2509-0119.

[16] S. Munteanu, V. Sudacevschi, V. Ababii, R. Braniste, A. Turcan, and V. Leashcenco. *Cognitive Distributed Computing System for Intelligent Agriculture.* International Journal of Progressive Sciences and Technologies. Vol. 24, No 2, January 2021, pp. 334-342. ISSN: 2509-0119.

[17] V. Ababii, V. Sudacevschi, R. Braniste, A. Nistiriuc, S. Munteanu, and O. Borozan. *Multi-Robot System Based on Swarm Intelligence for Optimal Solution Search.* In: The International Congress on Human-Computer Interaction, Optimization, and Robotic Applications, HORA-2020, June 26-28, 2020, Ankara, Turkey. pp. 269-273, Publisher: IEEE Catalog Number CFP20X32-ART, ISBN: 978-1-7281-9352-6, DOI: 10.1109/HORA49412.2020.9152926.

[18] V. Sudacevschi, S. Munteanu, V. Ababii, R. Braniste, O. Borozan, and V. Alexei. *Cognitive Computing System based on Distributed Knowledge*. In: Extended Abstracts of the 10th International Conference on Electronics,

Communications and Computing (ECCO-2019), 23-26 October 2019, Chisinau, Moldova, pp. 98, ISBN: 978-9975-108-84-3.

[19] V. Ababii, V. Sudacevschi, M. Osovschi, A. Turcan, A. Nistiriuc, D. Bordian, and S. Munteanu. *Distributed System for Real-Time Collective Computing.* In: Proceedings of the Fifth Conference of Mathematical Society of Moldova, IMCS-2019, September 28 – October 1, 2019, Chisinau, pp. 267-274. ISBN: 978-9975-68-378-4.

Victor Ababii[1], Viorica Sudacevschi[2], Silvia Munteanu[3], Victoria Alexei[4], Radu Melnic[5], Ana Turcan[6], Vadim Struna[7]

[1]Technical University of Moldova, Computer Sciences & Systems Engineering Department
E-mail: victor.ababii@calc.utm.md

[2] Technical University of Moldova, Computer Sciences & Systems Engineering Department
E-mail: viorica.sudacevschi@calc.utm.md

[3] Technical University of Moldova, Computer Sciences & Systems Engineering Department
E-mail: silvia.munteanu@calc.utm.md

[4] Technical University of Moldova, Computer Sciences & Systems Engineering Department
E-mail: victoria.lazu@ia.utm.md

[5] Technical University of Moldova
E-mail: radu.melnic@adm.utm.md

[6] Technical University of Moldova, Computer Sciences & Systems Engineering Department
E-mail: ana.turcan@fcim.utm.md

[7] Technical University of Moldova, Computer Sciences & Systems Engineering Department
E-mail: vadim.struna@iis.utm.md

The Plirophoria: The Missing Puzzle of the Ultimate Picture of the Universal Information

Veaceslav Albu

Abstract

The paper introduces the term Plirophoria to designate two kinds of the concept of information that modern science embraces nowadays. Such differentiation is of great importance as it permeates to make sense of information from one side, in the classical view, as a human artifact that is deeply theorized in information theory, and about evolutionary information that emerges by natural kinds in the Universe without any human assistance, driven by universal laws of nature. The four (but not all) necessary features of Plirophoria are defined. In addition, that differentiation has the potential to path the way towards a new theory of universal information by natural kinds that have the same degree of ontological objectivity for building the reality of the Universe, as matter and energy have.

Keywords: Plirophoria**,** information of natural kinds, evolutionarity, duality, objectivity, homeostaticity.

1. The Exordium

Norbert Wiener by his famous citation: "Information is information, not matter or energy. No materialism, which does not admit this, can survive at the present day" [1], pointed out the importance to recognize the sameness of the ontological status for information, matter, and energy (I/M/E) in nature. Modern science physicists take the concept of information as an artifact of interaction between matter and energy under the slogan "Information is the difference that makes the difference". The firm reductionist stand of modern theoretical physicists ignores N.

Wiener's legacy mentioned above. Nevertheless, not only theoretical physicists treat the concept of information as an artifact. Scientists from such fields as pancomputationalism, panpsyhism, and the theory of integrated information that specifically use information and computation as central concepts do the same, trying to build an integral picture of the Universe as a sort of computer.

Modern theoretical computer science, including AI, neural networks, and complex systems theory takes undoubtedly the anti-reductionist stand and continues to flourish by implementing the cybernetic systems of different orders towards building the informational society. The implemented concept of homeostasis gives modern cybernetics the grounds for an explanation of such properties of cybernetic systems as "self-subsisting, self-replicating, self-organizing, self-developing, self-assembling, self-regulatory, self-controlling, self-interaction, self-folding, self-learning", and this list can be prolonged. At the same time, the physicists often take as given those properties for matter or physical phenomenon in the Universe without any further grounded explanation. In this work, we will focus on formulating the necessary features and properties of the Universal concept of information, which can fit with the point of view on the information of the existing theories as an artifact of the human mind, as well as the main requirement of universal ontological objectivity and fundamentality of the concept of information alongside with matter and energy.

The goal of this work consists in constructing an enlarged version of the universal concept of information, which will contain the existing one as a part. Such enlarged concept of information should engulf the information by natural kinds evolved in the Universe from the very Bing-Bang's moment to nowadays evolution of the phenomenon of Life on Earth, independently of the fact of observation by any type of minds. At this point, it becomes clear that we are in need to have a clear differentiation of these objectively existing two kinds of information. In this work, we introduce the term "Plirophoria" to denominate the objectively existing evolutionary information by natural kinds. This is the antic Greek word used for the description of "information" in a universal sense and is not widely used in a nowadays-scientific publication. This will make it possible for this term not to be confused with information as

an artifact of the description of a difference in processes of interaction of matter or energy, which we will call information in Shannon's sense.

In the first section of this work, we provide additional evidence of the difference between the aforementioned two kinds of information, based on the Informational Catastrophe argument described in the paper [2]. We show on this basis the difference between operating and storing information of natural kinds and the information as an artifact. In the second section, we will describe four (but not all) necessary properties, which Plirophoria should possess in order to have the potential to build an explanatory theory that claims the sameness of the ontological status of Matter, Energy, and Information in the Universe.

2. The Digital Information Catastrophe

We use the above-mentioned Information Catastrophe argument with a clear purpose – to use it as the evidence that modern concept of information as an artifact and the proposed concept of Plirophoria do not represent the same thing. The Universe's Plirophoria cannot be reduced to the existing concept of information. It means that the concept of Plirophoria as universal information possesses the potential to enhance today's dominating notion of information and opens by that the possibility of embedding the Plirophoria into a modern scientific paradigm. We do not agree with the author of mentioned argument on many points, especially in its version of the matter/energy/information equivalence principle initially formulated by Albert Einstein [3] in 1905.

The argument about the information catastrophe was formulated with a clear aim to point to the possible coming future disaster of the era of digital information in human civilization caused by the emergence and evolution of cybernetic systems. The author argues that "… we will have 10**50 bits of information after 6000 years from now. This number of bits is very significant because it represents the approximate number of all atoms on Earth. … Even assuming that future technological progress brings the bit size down to sizes closer to the atom itself, this volume of digital information will take up more than the size of the planet, leading to what we define as the information catastrophe" [2].

From the above article, one can see that according to the presented figures, the flourishing of our informational society is under a big threat.

We can suppose that some future scientists will blame cybernetics for those consequences and as a computer scientist, the author cannot afford that as Norbert Wiener put it clearly in his above-presented citation. His genius got a clear intuition of what information in living organisms means and how it can be used for humans in constructing the "digital hammer" of the new information era. As principles of fluctuation of information in living and cybernetic systems coincide, by analogy with wave's theory, one can say that the digital information era in human history represents a resonance superposition of these two kinds of information, which drives the exponential evolution of the modern society. However, how producing of information by natural kinds works in such closed systems as our planet?

Let us present also some calculations. For the sake of simplicity, we will take as a bit of information a quantum event like changes or disruption of any chemical valence bond between two atoms or molecules. In our previous work, we claimed: "Taking into account that the average number of chemical reactions per cell per second equals to 10**8, the results of chemical reactions happening in all living cells on Earth represents 3.0 x10**38" [4]. Nevertheless, that calculation does not take into account the all-chemical reaction that takes place in other places on Earth. For example, the reaction of photosynthesis due to photons' wind from the Sun consists of about 10**17 photons per second per square centimeter. The absorption of a photon by an atom represents also a quantum event that can be taken as a bit of change of information. For all illuminated parts of Earth of 2.55x10**18 cm2, we have an estimated amount of 10**35 photons per second hitting the Earth.

Due to these facts, the total amount of information created and evolutionary stored by nature only during one billion years of evolution of life on Earth can be easily approximated as 3.0x10**38 multiplied by seconds per year 3.154x10**7 and by minimum one billion (1x10**9) years of evolution of life. Finally, we got the total number of bits of information created on Earth only by evolution of life equal to 0.946x10**55. According to the Informational Catastrophe argument, if the emergent information by natural kinds in living organisms possesses the same properties as digital information created by humans, the phenomenon of Life on Earth has to be extinct many millions of years

ago. Moreover, it is crucial for the sake of our paper to mention, that all the above amount of evolutionary information is created only by involving the 0.000008 percent [5] or 1x10**(-8) of the total mass of all atoms on Earth.

Conclusion. The Information (Digital) Catastrophe argument and presented calculation on emergent evolutionary information by natural kinds in living cells on Earth provides clearly the evidence, that information in living systems and in the Universe does possess additional properties that are not taken into account yet by the modern scientific paradigm. Norbert Winner tried to communicate to us that information is more than a digital bit or qubit. In this work, we will name several properties of universal evolutionary information by natural kinds, or Plirophoria, that can shed light on the ontological status of Plirophoria as evolutionary information in the Universe.

3. The necessary features and properties of Plirophoria

As cybernetic systems' main properties and components were observed in the functioning of living organisms, it is logically possible to presuppose that properties of the dynamically organized flux of information in such systems will mirror some properties of information emerging by natural kinds in living organisms. Below we provide the description of four properties of Plirophoria that are not associated with objective properties of information in Shannon's sense. However, that can be useful for us as hints to understand how Plirophoria is organized and how it functions in the Universe.

Evolutionarity. In any cybernetic system, all needed evolutionary information in the system is stored and available at any time when the system needs it (from the system's memory). We observe the same in our consciousness – when we recall some moments from the past, we can do it with a very high degree of reality. Some humans remember all past events in a very detailed way as a video does. Some astronomers describe our Universe as simultaneously existing events from billions of years ago up to a few days ago. Nevertheless, we know that they are not happening in one moment but represent the evolutionary history of the Universe. Therefore, evolutionarity of information by natural kinds, or to put it in another way, the instantaneous access to all evolutionary information of

the system represents common features of both natural informational systems and cybernetic systems as well. Hence, the existing definition and properties of information in modern science do not represent the way, how the Universe gets access to this information, as humans and computers do.

Objectivity. The emergent information by natural kinds in the Universe, described in the previous section, exists independently of any observer and plays a key objective role in its evolution. Modern science cannot deny that even some quantum physics principles are not explained yet by modern science. The existing scientific definition of information does not possess the potential to answer these paradoxes. All attempts of modern science to explain reality are based only on ontological objectivity of matter and energy, not involving information in that ontology. Information plays a descriptive role of resulted interaction of material particles without any information-based causal implications. Due to the lack of ontological objectivity of the concept of information, a new mysterianism flux appeared in modern science. The well-known scientists sustain that there are embedded limits of our knowledge and human brain, which we are not able to overcome and we will never know all the truth about how the Universe functions. In our opinion, the claim of objectivity of information by natural kinds that persists in the evolution of the Universe is crucial for the new concept of universal information, namely Plirophoria.

Duality. The modern science theories claim the experimentally demonstrated dualism of matter and energy at a quantum level. Particularly, photons and electrons possess such a wave – corpuscular quantum dualism. From one side, in the presented approach to information by natural kinds, we claim the ontological equivalence of matter, energy, and information. On the other side, matter and energy do possess the dualism property at a quantum level. From this, we logically conclude that Plirophoria, as the concept of universal information by natural kinds, should embrace the claim of possession of the same property – the quantum dualism of information.

Homeostaticity. The living systems and cybernetic systems have in common the homeostasis property in their functioning. Modern science also demonstrates that not only the living systems but also such complex

systems as atmosphere, stars, galaxies, and even galaxies clusters and black holes possess homeostatic or self-regulating properties. Therefore, we have to conclude that to build a viable concept of Plirophoria, we are to have in mind that this concept should explain how the Plirophoria could complement and sustain the functioning of a mechanism of the control center of such homeostatic systems. Moreover, such a concept needs to have the potential to answer the challenge of the possibility of existing of an extra dimension, as the control center for all cybernetic systems is situated in another dimension – dimension of human minds.

4. Conclusion

In this paper, we introduced the term Plirophoria with a clear purpose – to ground the path towards building the new theory of universal evolutionary information of natural kinds. We have a clear vision and scientific belief that the new concept of information that puts together the basics of functioning of our Universe's matter, energy, and information is missing in modern science's paradigm. The theory of Plirophoria that claims the sameness of ontological status of I/M/E will open the possibilities for formulating and updating scientific theories able to explain our reality with the highest degree of credibility and without need in a new mysterianism. The further development of the theory of Plirophoria, including additional properties of Plirophoria such as *fractality* and *imaginarity* (*or extra-dimensionality*) will be presented in forthcoming papers.

Acknowledgments. The research was supported by project 20.80009.5007.22. Intelligent information systems for solving ill-structured problems, processing knowledge and big data.

References

[1] N. Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*, MIT Press, 2nd revised ed., 1961, 352 p.

[2] Melvin M. Vopson. *The information catastrophe*. AIP Advances **10**, 085014 (2020). DOI: 10.1063/5.0019941, Published Online: 11 August 2020.

[3] A. Einstein. *Does the inertia of a body depend upon its energy content?* Annalen der Physik, **18,** 1905, pp. 639-641 (https://www.astro.puc.cl/~rparra/tools/PAPERS/e_mc2.pdf).

[4] V. Albu. *A Road Map for an Informational Ontology of Information/Energy/Matter' Homeostatic Unity Inspired by Information Catastrophe*, In: Proceedings of the Workshop on Intelligent Information Systems WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova, pp. 15-24.

[5] *The Biosphere.* New World Encyclopedia, https://www.newworldencyclopedia.org/entry/biosphere (accessed on 26 of July 2021)

Veaceslav Albu

Vladimir Andrunachievici Institute of Mathematics and Computer Science
E-mail: vaalbu@googlemail.com

Development of computing infrastructure for support of Open Science in Moldova

Petru Bogatencov, Grigore Secrieru,
Boris Hîncu, Nichita Degteariov

Abstract

In the paper, there are considered the e-Infrastructures which enable instruments and provide facilities, resources, and services that are used by the research communities to conduct research and foster innovation in their fields. The European Open Science Cloud (EOSC) will be the open and trusted virtual environment which will allow European researchers to store, share and reuse research data across borders and disciplines. Elaboration of the distributed computing infrastructure to support the Open Science initiative in Moldova has an important role for the research community in using the performances of the EOSC ecosystem.

Keywords: computing infrastructure, research infrastructures, e-Infrastructures, EOSC ecosystem, platforms, and tools for Open Science.

1. Introduction

The exponential growth of information and the availability of digital technologies have generated a new approach to scientific research - Open Science, which is based on accessibility, collaboration, e-infrastructures, and new ways of disseminating knowledge. The following key circumstances are associated with the development of the infrastructure for Open Science support:

- Most of the existing infrastructure of Open Science, in addition to open access to publications, can be considered to be at the beginning of the path;

- We need a stable, reliable, and evolving infrastructure that will support the creation of more research data, with the possibility of their reuse, as well as digital research tools that are efficient and easy to use,
- Some elements of the global open science infrastructure require collective efforts for their creation and management.

These circumstances aim to achieve a consensus on ensuring equitable access to the advantages of Open Science and participation in its development.

The main support for the development of Open Science is provided by the EOSC ecosystem [1], which will be developed in the period 2021-2027 within the Horizon Europe research-innovation program. EOSC constitutes a major ambition in the European Open Science policy, being a federated ecosystem of research infrastructures, e-infrastructures, and services that allow the scientific community to share and process publicly funded research results and data across borders and scientific domains.

The National Open Science Cloud Initiatives (NOSCI) are important pillars of this effort [2]. For ensuring the sustainability of this flagship initiative for Europe, it has to be built upon solid governance and organizational framework on the national level.

In order to stimulate the participation of countries from different regions of Europe in the promotion, expansion, and use of the resources of the EOSC ecosystem in progress are European projects, including the project “NI4OS Europe – National Initiatives for Open Science in Europe”. The NI4OS-Europe project proposes a modular workflow for the integration of national resources and services into EOSC and the setup of NOSCI. RENAM Association, as a participant in the NI4OS-Europe, initiated, according to the project requirements, the process of establishing the national initiative on the Cloud for Open Science – MD-NOSCI, developed "NOSCI Establishment Roadmap for the Republic of Moldova", presented to the Ministry of Education and Research. In this way, RENAM provides the Moldovan research community with information and access to up-to-date resources, representing the main point of contact with international initiatives in the field of open science. NOSCIs in the Member States and partner countries have an important

role to play in using performance and facilitating EOSC governance. Fig. 1 expresses an overview of the EOSC ecosystem, presented in [1].

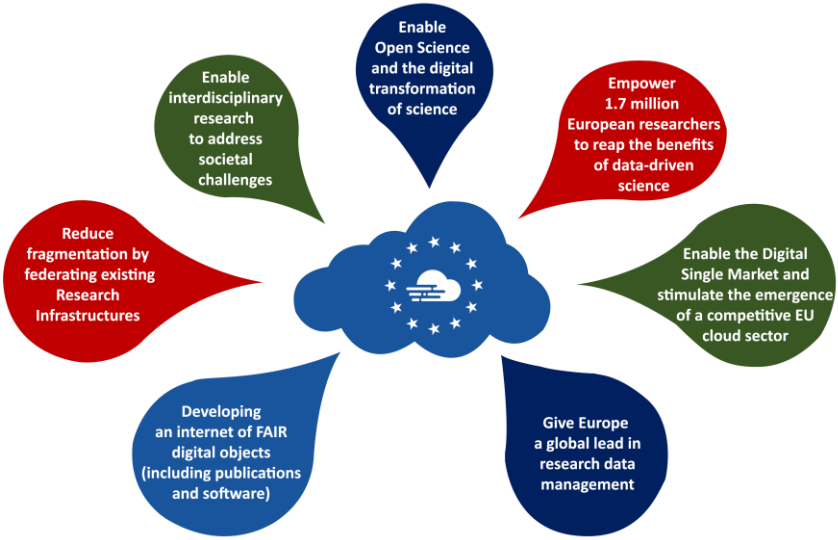

Figure 1. EOSC vision.

NOSCI can be seen as a coalition of national organizations that have a prominent role and interest in the European Open Science Cloud. The main goal of NOSCI is to promote synergies at the national level and to optimize their participation in European and global challenges in the field of open science, the development of the EOSC ecosystem that includes as an example the elaboration of the software infrastructure for the cloud-like high-performance computing oriented towards classes of problems in the field of mathematical modeling of decision-making processes.

The details of the NOSCI establishment objectives and the necessary activities can be provided in the Memorandum of Understanding that will be the basis of this initiative.

The State University of Moldova (SUM) and the RENAM share the goal of developing and providing an innovative federated Cloud environment with the efficient use of hard and soft resources available to promote the concept of Open Science, development of digital technologies

in research and education (R&E) and solving problems that require high-performance computing (HPC) and generating large volumes of data.

For the interconnection of universities and research institutes, the RENAM-GEANT networking platform is used at speeds up to 10 Gbps and more based on the national fiber optic backbone of NREN RENAM. These actions are aimed at creating the necessary conditions for the academic and research community to use the performance of the EOSC ecosystem.

2. Development of distributed computing infrastructure

Nowadays, almost any organization has its own IT structure, which unites servers and computers of employees into a common network, or can rent cloud services and refuse from owning physical servers. In describing cloud solutions, there are often three cloud service models that have the following abbreviations: IaaS, PaaS, and SaaS.

Infrastructure as a Service (IaaS) is a solution for hosting infrastructure in the cloud by renting cloud servers. This means creating an infrastructure in the cloud with the required configuration of computing resources in the form of CPU, RAM, HDD, which the customer uses at his own discretion. It is an alternative to investing in expensive hardware and networking equipment. IaaS allows you to organize a pool of virtual machines, prepare remote workstations, data storage, etc., providing quick access to reliable, secure, flexible, and scalable infrastructure.

The IaaS configuration model for computing resources, hardware, and network equipment includes:

1. Virtual servers (VPS / VDS) on which you can install various software products. The provider can offer servers with operating systems so that you can quickly deploy the necessary applications to them.
2. Network links that allow virtual servers to communicate with each other, external servers owned by the customer, and the Internet. These include:
    - server availability for each other and the external network, routing of server network connections;
    - load balancing, which prevents server overload by distributing incoming traffic between the pool of servers;

- VPN – technology for encrypting data transmitted by a company between the cloud and its physical data center;

3. User access control. For example, you can restrict access to individual virtual machines or allow viewing of data, but prohibit making changes to them.
4. Cloud storage for storing files, data, or backups. They differ from ordinary cloud drives, which individual users deal with, with almost unlimited storage capacity and high speed of data access.
5. Backup services and resilience disasters that insure infrastructure against falls and data loss in the event of failure of its individual nodes.

The Platform as a Service (PaaS) and the main differences from the IaaS model. In the case of PaaS, the customer is provided with certain tools such as a database management system, a big data processing environment, which need to be customized to the needs of the organization, but do not need to be built from scratch. At the same time, there is no access to the operating system and the settings of virtual servers that underlie PaaS, there is only access to the interfaces of the platform itself. In the case of IaaS, we only get disk space and must choose an open research data management system (ORDMS), install and configure it, ensure data protection and backup. In PaaS, the ORDMS is already installed, you just need to configure it for yourself and manage your data.

The Software as a Service (SaaS) model is a fully configured, out-of-the-box software environment that performs specific functions. The software itself is in the cloud and is accessed via a network, and the software environment runs on the capacities of virtual servers. Most of the services on the Internet can serve as examples of SaaS: email, task schedulers, web builders for creating sites, open research data repositories, and other cloud applications for solving specific problems.

SUM, Vladimir Andrunachievici Institute of Mathematics and Computer Science (IMCS), and RENAM jointly realize the goal of developing and providing resources of the innovative federated Cloud environment that ensures the efficient use of available hardware and software resources to promote the concept of Open Science, development

of digital technologies for research and education (R&E) and solving problems that require high-performance computing resources and processing of large amounts of data.

The key points for the creation and development of a distributed infrastructure to support open science were facilitated by the participation of SUM, RENAM, and IMCS in international and national projects that allowed developing and modernizing the computing hardware and network equipment: SUM, RENAM, and IMCS computing clusters were upgraded, new high-performance servers for compute nodes and storage elements were purchased; 10 Gbps optical network deployed between the main computing locations; IMCS and central RENAM nodes were modernized – modern uninterruptible power supply systems for server equipment and industrial air cooling systems were installed.

Current activities focused on deployment and development of HPC cloud infrastructure by implementing SaaS, PaaS, and IaaS service models for the joint use of parallel computing clusters of SUM, IMCS, and RENAM integrated into a single distributed Cloud environment. Thus, a cloud service will be created that is focused on performing various types of tasks, hosting ORDMS, and is suitable for deploying open science tools and services.

Until now Cloud IaaS based on OpenStack [3] is running on the IMCS-RENAM computing infrastructure and has been available for scientists for several years. The system is successfully used by researchers in several resource-intensive projects that use Machine-Learning technologies to train neural networks in text recognition, language processing, etc. The existing system has some disadvantages due to the limited number of available resources and lack of a high throughput network interconnection between nodes and storage elements [4]. The creation of the new joint SUM-RENAM-IMCS Cloud infrastructure is aimed at solving these known problems.

3. Expected results and future plans

At the first stage, it is planned to create a multi-zone IaaS Cloud infrastructure that combines the resources of IMCS, SUM, and RENAM into a distributed computing network for processing scientific data,

performing scientific calculations, as well as storing and archiving results (see Fig. 2).

PaaS and SaaS solutions will be deployed on the created computing infrastructure: tools for processing and analyzing scientific data, such as Juniper Notebook, Python, Elastic Search, and others; open repositories for scientific data collection will be installed. The tools and platforms proposed for deployment will be integrated (onboarded) into the EOSC Portal.

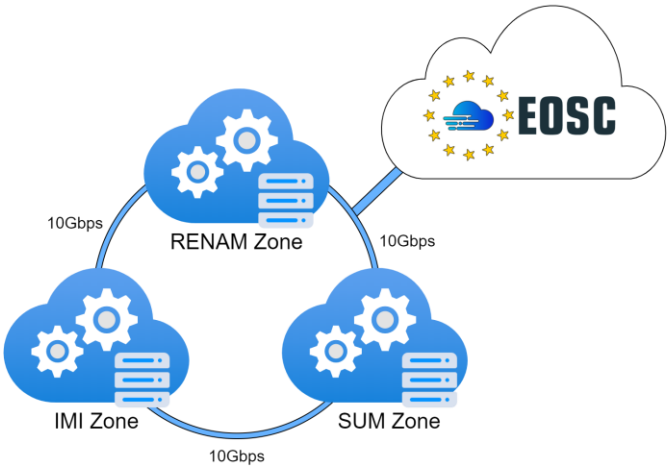

Figure 2. Multi-zone Cloud: IMI – RENAM – SUM.

Integration of new servers will increase the computing power, and the combination of computing power and data storage systems of the three organizations will allow more flexible and efficient use of computing resources for solving large-scale problems. Parameters of the newly created distributed computing infrastructure:

- IMCS Compute servers: Dell R540 servers with total of 32 CPU cores, 128GB RAM, Dell R740 server with 12 TB of RAID of storage;
- RENAM Compute servers: Dell R730 servers with total of 40 CPU cores, 256 GB RAM, Dell R 740 server with 12 TB of RAID storage;
- SUM Compute servers: HP ProLiant DL140 G3 with total of 28 CPU cores, 82 GB RAM, 2TB of RAID Storage.

An analysis of the hardware and software resources of the parallel clusters of SUM, RENAM, and IMCS will be carried out with the aim of coordinated and safe use of joint resources. Optimal configurations of hardware and software resources of available clusters and servers will be elaborated and realized for effective use of the available resources in the Cloud infrastructure.

Distributed computing infrastructure will use the adaptive execution framework that can be adapted and tested for the solution of different complex applications. The research team from SUM developed several applications that require resources of multiprocessor clusters and distributed computing infrastructure. One practical example of the application for solving complex decision-making problems – elaborating for porting on the created computing infrastructure – is described below.

3.1 The parallel algorithm for determining Nash equilibrium profiles in modeling decision-making processes.

As mentioned above the developed Cloud HPC computing system can be used for operating data storage and analysis platforms, executing various scientific applications, including modeling decision-making processes in situations of risk, conflict, and informational impact, using the mathematical apparatus of game theory. In this paragraph, we will describe and analyze how to test the Cloud HPC system by running parallel programs. As an example, the parallel algorithm for determining Nash equilibrium situations in modeling decision-making processes is proposed.

Contemporary decision-making problems are very complex and require the processing of a very large volume of data. Thus, for the mathematical modeling of these processes, it is necessary to take into account the big data problems. Big data is a huge amount of data that is beyond the processing capacity to manage and analyze the data in a specific time interval. The data is too big to be stored and processed by a single machine. In many large-scale solutions, data is divided into partitions that can be managed and accessed separately. In order to solve such problems in real-time, parallel algorithms are built and then implemented on various types of parallel computing systems. For parallel data processing, we must use the ways of dividing, partitioning (sharing), and distributing data. Different paradigms and programming models can

be used for the soft implementation of parallel algorithms on distributed memory computing systems (parallel clusters, cloud computing systems). From the multitude of parallel programming models here we will present how to implement software on HPC clusters of the model based on MPI functions.

For modeling the decision-making problems we will consider the bimatrix game in the following strategic form $\Gamma = \langle I, J, A, B \rangle$, where *I={1,2,..,n}* is the line index set (the set of strategies of the player 1), *J={1,2,..,m}* is the column index set (the set of strategies of the player 2) and $A = \|a_{ij}\|_{\substack{i \in I \\ j \in J}}$, $B = \|b_{ij}\|_{\substack{i \in I \\ j \in J}}$ are the payoff matrices of player 1 and player 2, respectively. We denote by *NE[Γ]* the set of all equilibrium profiles in the game *Γ*. Thus, the Nash equilibrium profile is the pair of indices (i^*, j^*), for which the following system of inequalities is verified $(i^*, j^*) \Leftrightarrow \begin{cases} a_{i^* j^*} \geq a_{ij^*} \ \forall i \in I, \\ b_{i^* j^*} \geq b_{i^* j} \ \forall j \in J. \end{cases}$ Based on this definition, it is easy to develop the parallel algorithm for determining the Nash equilibrium profiles in bimatrix games.

The structure of the parallel algorithm is determined by the parallelization mode at the data level. That is, the following ways of dividing and distributing matrices A and B can be used:

- Matrices are divided into rectangular submatrices of any size. In this case, the way of constructing equilibrium profiles for the game with the initial matrices is very complicated;
- Matrices are divided into line-type submatrices or column-type submatrices. In this case, building the equilibrium profiles for the initial game is quite simple.

We will mathematically describe the parallel algorithm for determining Nash equilibrium profiless in pure strategies for the bimatrix game defined above. We will assume that matrix A is divided into column-type submatrices and matrix *B* is divided into row-type submatrices. So, we're going to get a series of submatrices $SubA^t = \|a_{ij}\|_{\substack{i \in I \\ }}^{j \in J_k}$ and $SubB^t = \|b_{ij}\|_{i \in I_k}^{j \in J}$, where $I_k = \{i_k, i_{k+1}, ..., i_{k+p}\}$ and

$J_k = \{j_k, j_{k+1}, ..., j_{k+p}\}$. $SubA^t$ is a submatrix that consists of p columns of matrix A starting with column number k and is "distributed" to the process with the rank t. Similarly, SubBt is a submatrix that consists of p lines of the matrix B starting with the line k and is also distributed to the process with the rank t. Using the sequential algorithm described above, the process with the rank t will determine a graph of the point-to-set application $i^*(j_k) = Arg\max_{i \in I} a_{ij_k}$. for any $j_k \in J_k$. Similarly, the process with the rank t will determine a graph of the point-to-set application $j^*(i_k) = Arg\max_{j \in J} b_{i_k j}$. for any $i_k \in I_k$. Finally, the process with rank t will determine $LineGr^t = \bigcup_k gr_k i^*, ColGr^t = \bigcup_k gr_k j^*$. So the Nash equilibrium profiles will be the intersections of the set $\left(\bigcup_t LineGr^t\right)$ and $\left(\bigcup_t ColGr^t\right)$.

Using the MPI parallel programming model [5] on the parallel computing system with distributed memory, a parallel program was developed and tested on control examples to determine Nash equilibrium profiles in bimatrix games. The results of the calculations are presented in Table 1:

Table 1. Computation time for determining solutions in two-matrix games.

Dimensions of the matrices	Runing time (seconds)			
	16 processors	28 processors	32 processors	48 processors
n=30000 m=30000	11.665889	11.187384	14.855256	14.361466
n=35000 m=35000	18.748501	12.359752	13.841478	10.827413
n=40000 m=40000	17.021988	12.191731	13.195590	11.202775
n=45000 m=45000	33.997708	14.808254	13.028965	12.560618
n=48000 m=48000	69.756550	20.810968	18.946320	15.717546

n=49000 m=49000	110.552440	32.880968	16.161673	15.140105
n=49500 m=49500	186.125029	40.321999	22.643520	17.640198
n=49900 m=49900	eror	55.035883	17.819622	19.017421
n=50000 m=50000	eror	46.226555	16.388920	11.839626
n=55000 m=55000	eror	eror	57.292023	38.964008
n=60000 m=60000	eror	eror	eror	156.085372

We have to mention that the developed application we plan to use for testing efficiency of the developed distributed HPC Cloud system, i.e. to use it as a benchmarking program.

4. Conclusion

At the European level, the main support for the development of open science is provided by the EOSC ecosystem, which will be developed in the period 2021-2027 within the Horizon Europe research-innovation program. The possible way to activate participation of the national research community in the European Open Science initiatives is to deploy integrated to EOSC national e-Infrastructures and services. The proposed approach of creation of joint infrastructure and resources for support of Open Science at the national level is one important pave in this direction.

As a principal organizational mechanism that will support the EOSC integration activities is the establishment of the National Initiative MD-NOSCI in Moldova to comply with the European and global challenges in the field of open science. National initiatives in the EU Member States and partner countries play an important role for the academic and research communities in the development and use of open research data and services, as well as in facilitating the governance of the EOSC.

Acknowledgments. This work was supported by the National Agency for Science and Development (grant no. 20.80009.5007.22 and grant No.

20.80009.5007.13) and by the European Commission, project H2020 NI4OS-Europe (grant No. 857645).

References

[1] *EOSC* : https://www.eosc.eu/

[2] *NI4OS* : https://ni4os.eu/

[3] *Cloud IMI-RENAM*: https://cloud.renam.md/

[4] Nichita Degteariov, Petru Bogatencov, Nicolai Iliuha, and Grigore Secrieru. *Upgrading Cloud Infrastructure for Research Activities Support*. In: Workshop on Intelligent Information Systems (WIIS2020) Proceedings, Chisinau, IMI, 2020, pp. 69-74.

[5] B. Hîncu and E. Calmîş. *Modele de programare paralelă pe clustere. Partea I. Programare* MPI. Note de curs. Chişinău: CEP USM, 2016, 129 p.

Petru Bogatencov[1,3], Grigore Secrieru[1,3], Boris Hâncu[2], Nichita Degteariov[3]

[1]PhD/Vladimir Andrunachievici Institute of Mathematics and Computer Science. Chisinau, Moldova.
E-mail: bogatencov@renam.md, secrieru@renam.md

[2]PhD/State University of Moldova. Chisinau, Moldova.
E-mail: boris.hancu@gmail.com

[3]RENAM Association. Chisinau, Moldova.
E-mail: nichita.degteariov@renam.md

Proceedings of Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

Hamilton full favoring apportionments

Ion Bolun

Abstract

Aspects of full favoring of large beneficiaries in apportionments using Hamilton method are discussed. In this aim, the requirements of full favoring apportionments compliance with Hamilton method's solutions were defined. Subsequently, the A_{HL} algorithm for determining the Hamilton apportionments which fully favor large beneficiaries is described. Using this algorithm, calculations for two examples were performed. The obtained results confirm the opportunity of using the A_{HL} algorithm for the generation of Hamilton full favoring apportionments.

Keywords: algorithm, apportionment problem, disproportionality, Hamilton method, favoring of large beneficiaries.

1 Introduction

Often it is necessary to distribute a given number M of discrete entities of the same kind among n beneficiaries, in proportion to a numerical characteristic assigned to each of them V_i, $i = \overline{1,n}$. This is known as proportional apportionment (APP) problem [1, 2]. The integer character of this problem usually causes a certain disproportion of the apportionment $\{x_i,\ i = \overline{1,n}\}$ [1, 3], some beneficiaries being favored at the expense of the others. Such favoring leads to the increase of disproportionality of the apportionment. Therefore, reducing the favoring in question is one of the basic requirements when is choosing the APP method to be applied for apportionments.

In this aim, it is needed to estimate this property quantitatively. One approach is proposed in [5]. Another, the "total (full) favoring", is

examined in [6]. In [6], it was shown that the frequency of full favoring in apportionments, for the widely used Hamilton [3], Sainte-Laguë [3], d'Hondt [3] and Huntington-Hill [4] methods, is strongly decreasing on n, becoming approx. 0 at $n \geq$ 7-10. Aspects of the guaranteed generation of Hamilton apportionments, which fully favor large beneficiaries (with higher V_i value) at larger values of n, are examined in this paper.

2 Essence of full favoring in apportionments

The notion of "total (full) favoring" of beneficiaries was introduced in [6] based on the definition of favoring of large or of small beneficiaries by an APP method given in [1] (see Definition (1)).

Definition 1. *In an apportionment, large beneficiaries are fully favored if*

$$\frac{x_i}{V_i} > \frac{x_j}{V_j}, \tag{1}$$

whenever $x_i > x_j$, *where* $(i, j) \in \{1, 2, 3, \ldots, n\}$ *[6].*

Usually, in one and the same apportionment some large and some small beneficiaries can be favored and, nevertheless, mainly large or, on the contrary, mainly small beneficiaries are favored. Therefore, in [5] it is proposed to use two different notions: "favoring" of large or of small beneficiaries and "full favoring" of large or of small beneficiaries, the second being a particular case of the first one. The compliance of an apportionment with requirements (1) is referred to "full favoring" of large beneficiaries. The larger notion of "favoring" is used when in an apportionment large beneficiaries are predominantly favored or, on the contrary, the small ones in sense of [5].

In order to identify whether apportionments that fully favor large beneficiaries can be obtained when applying the Hamilton APP method, it is necessary to know the compliance conditions of this method with requirements (1).

3 Compliance of Hamilton apportionments with requirements (1)

The required apportionments have to be Hamilton's ones and, at the same time, be compliant with requirements (1). The conditions for the compliance of an apportionment with the solution obtained by Hamilton method (Hamilton apportionment) are defined by Statement 1. First, let: $Q = V/M$; $V_i = a_iQ + \Delta V_i > 0$, $i = \overline{1,n}$; $\Delta M = (\Delta V_1 + \Delta V_2 + \Delta V_3 + \ldots + \Delta V_n)/Q$, $1 \leq l \leq n-1$ and $x_i > x_{i+1}$, $i = \overline{1,n-1}$. Of course, occur $0 \leq \Delta V_i < Q$, $i = \overline{1,n}$.

Statement 1. *The necessary conditions for the compliance of an apportionment* $\{x_i, i = \overline{1,n}\}$*, which fully favors large beneficiaries, with the solution obtained by Hamilton method are*

$$\Delta V_i > \Delta V_k, i = \overline{1,l}, k = \overline{l+1,n}. \tag{2}$$

Indeed, the Hamilton method apportionment rule states [3] that in addition to the already apportioned a_i entities, $i = \overline{1,n}$, the remained unapportioned $\Delta M = l$ entities should be apportioned by one to the first beneficiaries with the largest ΔV_j value. So, taking into account that $x_i > x_{i+1}, i = \overline{1,n-1}$, the relations $x_i = a_i + 1, i = \overline{1,l}$ and $x_i = a_i, i = \overline{l+1,n}$ should take place when favoring large beneficiaries, and that can be only if occurs (2). ♦

It should be mentioned that Statement 1 establishes relationships between beneficiaries of two groups, $\{i = \overline{1,l}$ and $i = \overline{l+1,n}\}$, but not between beneficiaries within each of these groups if $n > 2$, needed to be established when analyzing the full favoring of large beneficiaries according to requirements (1).

It is well known that overall, on an infinity of apportionments, Hamilton method doesn't favor beneficiaries [1, 3]. But it can be particular Hamilton apportionments which fully favor large beneficiaries. The respective requirements are defined by Statement 2.

Statement 2. *If* $n > 2$ *and* $l = \Delta M$*, the conditions for the compliance of a Hamilton apportionment* $\{x_i, i = \overline{1,n}\}$ *with the requirement (1) of*

full favoring of large beneficiaries, in addition to the (2) ones, are

$$\Delta V_i < \frac{a_i}{a_{i+1}} \Delta V_{i+1}, \tag{3}$$

where $i = \overline{l+1, n-1}$ *if* $a_n > 0$ *and* $i = \overline{l+1, n-2}$ *if* $a_n = 0$ *for* $1 = l < n-1$ *(Case L1),*

$$\Delta V_i < \frac{\Delta V_{i+1}(a_i + 1) - Q(a_i - a_{i+1})}{a_{i+1} + 1}, i = \overline{1, l-1} \tag{4}$$

for $1 < l = n-1$ *(Case L2) and both, (3) and (4), for* $1 < l < n-1$ *(Case L3).*

Indeed, one has $0 \le \Delta V_i < Q, i = \overline{1, n}$. Let's begin with **Case L3**, divided into three subcases:

L3a) $x_i = a_i + 1, x_k = a_k + 1, i = \overline{1, l-1}, k = \overline{i+1, l}$;

L3b) $x_i = a_i + 1, x_k = a_k, i = \overline{1, l}, k = \overline{l+1, n}$;

L3c) $x_i = a_i, x_k = a_k, i = \overline{l+1, n-1}, k = \overline{i+1, n}$.

In **Subcase L3a**, according to (1) it should be $x_i/V_i > x_k/V_k$, that is $(a_i+1)/(a_iQ+\Delta V_i) > (a_k+1)/(a_kQ+\Delta V_k), i = \overline{1, l-1}, k = \overline{i+1, l}$, from where one has

$$\Delta V_i < \frac{\Delta V_k(a_i + 1) - Q(a_i - a_k)}{a_k + 1}, i = \overline{1, l-1}, k = \overline{i+1, l}. \tag{5}$$

Let's show that requirements (5) are transitive. From (5), for $k = i + 1$ one has

$$\Delta V_i < \frac{\Delta V_{i+1}(a_i + 1) - Q(a_i - a_{i+1})}{a_{i+1} + 1}, i = \overline{1, l-1} \tag{6}$$

and, respectively,

$$\Delta V_{i+1} < \frac{\Delta V_{i+2}(a_{i+1} + 1) - Q(a_{i+1} - a_{i+2})}{a_{i+2} + 1}, i = \overline{1, l-2}. \tag{7}$$

Taking into account (7), requirement (6) can be transformed as follows

$$\Delta V_i < \frac{(a_i+1)\frac{\Delta V_{i+2}(a_{i+1}+1)-Q(a_{i+1}-a_{i+2})}{a_{i+2}+1} - Q(a_i-a_{i+1})}{a_{i+1}+1}$$
$$= \frac{\Delta V_{i+2}(a_i+1) - \frac{Q(a_{i+1}-a_{i+2})}{a_{i+1}+1}(a_i+1) - \frac{Q(a_i-a_{+1})}{a_{i+1}+1}(a_{i+2}+1)}{a_{i+2}+1}$$
$$= \frac{\Delta V_{i+2}(a_i+1) - Q(a_i-a_{i+2})}{a_{i+2}+1}, i = \overline{1, l-2}. \quad (8)$$

So, if relations (6) and (7) take place, then relation (8) occurs, too. The same way, one can show that occurs

$$\Delta V_i < \frac{\Delta V_{i+j}(a_i+1) - Q(a_i-a_{i+j})}{a_{i+j}+1}, i = \overline{1, l-1}, j = \overline{1, l-i}. \quad (9)$$

Thus, relations (5) are transitive and can be replaced by the (4) ones.▼

In **Subcase L3b**, according to (1), it should be $x_i/V_i > x_k/V_k$, that is $(a_i+1)/(a_iQ+\Delta V_i) > a_k/(a_kQ+\Delta V_k), i = \overline{1,l}, k = \overline{l+1,n}$, from where one has $\Delta V_k(a_i+1) > a_k(\Delta V_i - Q)$. Because of $0 \le \Delta V_i < Q$ and $\Delta V_k(a_i+1) \ge 0$, the requirements $\Delta V_k(a_i+1) > a_k(\Delta V_i - Q), i = \overline{1,l}, k = \overline{l+1,n}$ always take place, that's why Subcase L3b is not specified in Statement 2.▼

In **Subcase L3c**, according to (1), it should be $x_i/V_i > x_k/V_k$, that is $a_i/(a_iQ+\Delta V_i) > a_k/(a_kQ+\Delta V_k), i = \overline{l+1,n-1}, k = \overline{i+1,n}$, from where one has

$$\Delta V_i < \frac{a_i}{a_k}\Delta V_k, i = \overline{l+1, n-1}, k = \overline{i+1, n} \quad (10)$$

if $a_n > 0$ and $\Delta V_n a_i > \Delta V_i a_n = 0$ if $a_n = 0, i = \overline{l+1, n-1}$. The last inequality always takes place, therefore it is not included in (3).

It is easy to show that requirements (10) are transitive. From (10), one has $\Delta V_i < \frac{a_i}{a_{i+1}}\Delta V_{i+1}$ and $\Delta V_{i+1} < \frac{a_{i+1}}{a_{i+2}}\Delta V_{i+2}$, from where $\Delta V_i < \frac{a_i}{a_{i+1}}\frac{a_{i+1}}{a_{i+2}}\Delta V_{i+2} = \frac{a_i}{a_{i+2}}\Delta V_{i+2}$. In the same way one can show that

relations $\Delta V_i < \frac{a_i}{a_{i+j}}\Delta V_{i+j}, i = \overline{l+1, n-1}, j = \overline{1, n-i}$ occur. Thus, relations (10) are transitive and can be replaced by the (3) ones.▼

The proof for **Cases L1** and **L2**, taking into account proofs for Subcases L3a and L3c, are trivial. ♦

When generating apportionments that fully favor large beneficiaries, the inequalities

$$\Delta V_i > \frac{a_i}{a_{i-1}}\Delta V_{i-1}, i = \overline{l+2, n}, l = \overline{1, n-2}, \tag{11}$$

$$\Delta V_i > \frac{\Delta V_{i-1}(a_i + 1) + Q(a_{i-1} - a_i)}{a_{i-1} + 1}, i = \overline{2, l}, \tag{12}$$

equivalent to the (3) and (4) ones, are also useful.

4 Generating Hamilton apportionments

Based on Statements 1 and 2, the **A_{HL} algorithm** for the generation of Hamilton apportionments which fully favor large beneficiaries was elaborated. According to (11), the lower the value of ΔV_{l+1}, the lower the values of $\Delta V_i, i = \overline{l+2, n}$. Similarly, according to (12), the lower the value of ΔV_1, the lower the values of $\Delta V_i, i = \overline{2, l}$. Taking into account these observations, in Figure 1 the basic conceptual steps of the A_{HL} algorithm are shown, considering $V > M$ and that the value of ΔM is known.

At Steps 3 and 4 of the A_{HL} algorithm, to $\Delta V_i > 0, i = \overline{1, n}$ minimal possible values are allocated: at Step 3 – to $\Delta V_i, i = \overline{l+1, n}$ according to requirement (11) and beginning with the value of $\Delta V_{l+1} > 0$; at Step 4 – to $\Delta V_i, i = \overline{1, l}$ according to requirement (12) and beginning with the value of $\Delta V_1 > z = max\{\Delta V_{l+1}, \Delta V_{l+2}, \Delta V_{l+3}, \ldots, \Delta V_n\}$ because of requirement (2). If after these allocations one has $\Delta M > l$, that is $\Delta V > \Delta U$, where $\Delta U = \Delta MQ$, then the solution doesn't exist.

On the contrary, if $\Delta V < \Delta U$, then one has to increase ΔV aiming to reach $\Delta V = \Delta U$. Because of requirement (2), it is relevant to increase first, maximal possible, the values of $\Delta V_i, i = \overline{1, l}$ beginning with $\Delta V_l < Q$. This is done at Step 5 according to requirement (4).

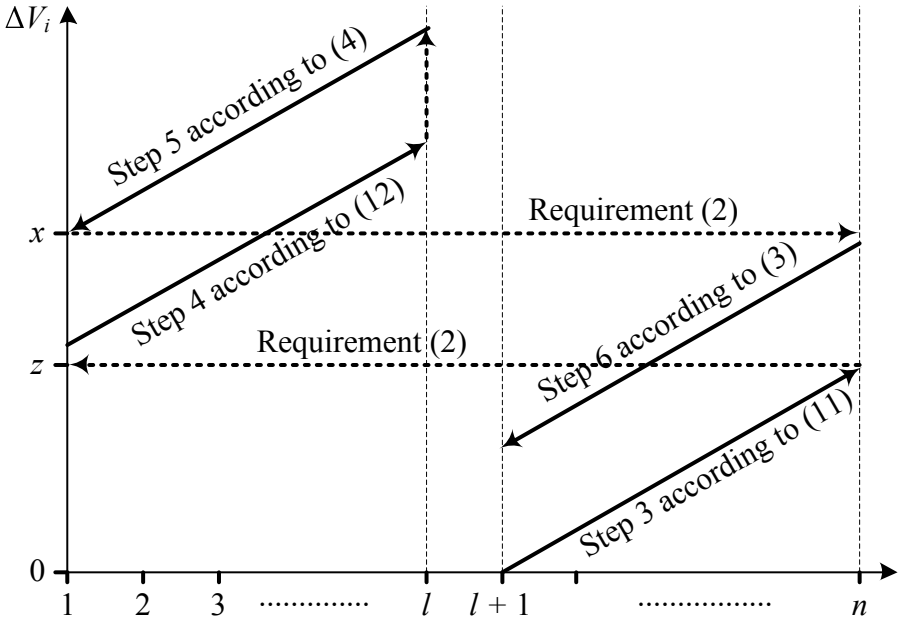

Figure 1. Basic steps of the A_{HL} algorithm.

But if the equality $\Delta V = \Delta U$ is not achieved at this step, then the last possibility to increase the ΔV value is the increase of $\Delta V_i, i = \overline{l+1, n}$ values beginning with $\Delta V_n < x = min\{\Delta V_i, i = \overline{1, l}\}$ because of requirement (2). This is done at Step 6 according to requirement (3).

It should be mentioned that in Figure 1 a continuous arrow doesn't reflect the relation between the values of ΔV_i and ΔV_{i-1}. It reflects the relation between ΔV_i and the respective function of:

1) ΔV_{i-1} (at Steps 3 and 4), that is $\Delta V_i > f_1(\Delta V_{i-1})$ according to requirement (11) and, respectively, the (12) one;

2) ΔV_{i+1} (at Steps 5 and 6), that is $\Delta V_i < f_2(\Delta V_{i+1})$ according to requirement (4) and, respectively, the (3) one.

The A_{HL} algorithm in details is described below.

1. Initial data are: $V, n, 1 \leq l \leq n-1, 1 \leq g \leq \lceil Q/n \rceil$ and $x_i > x_{i+1}, i = \overline{1, n-1}$.

2. $M := x_1+x_2+x_3+\ldots+x_n, Q := V/M, \Delta U := Ql; a_i := x_i-1, i = \overline{1, l}, a_i := x_i, i = \overline{l+1, n}$.

3. Based on (11), determining the preliminary, minimal possible, values of sizes $\Delta V_i \geq 0, i = \overline{l+1, n}$.
   3.1. $i := l+1, \Delta V_i := \lfloor Qa_i \rfloor + 1 - Qa_i$. If $i = n$, then go to Step 4.
   3.2. $i := i+1, \Delta V_i := \lfloor Qa_i + \Delta V_{i-1} a_i / a_{i-1} \rfloor + g - Qa_i$. If $\Delta V_i \geq Q$, then the solution doesn't exist. Stop.
   3.3. If $i < n$, then go to Step 3.2.

4. Based on (12), determining the preliminary, minimal possible, values of sizes $\Delta V_i > 0, i = \overline{1, l}$.
   4.1. $z := max\{\Delta V_{l+1}, \Delta V_{l+2}, \Delta V_{l+3}, \ldots, \Delta V_n\}; \Delta V := \Delta V_{l+1} + \Delta V_{l+2} + \Delta V_{l+3} + \ldots + \Delta V_n$.
   4.2. $i := 1$. $\Delta V_i := \lfloor Qa_i + z \rfloor + g - Qa_i$. If $\Delta V_i \geq Q$, then the solution doesn't exist. Stop.
   4.3. If $i = l$, then go to Step 5.
   4.4. $i := i+1$. $\Delta V_i := \lfloor Qa_i + [\Delta V_{i-1}(a_i+1) + Q(a_{i-1}-a_i)]/(a_{i-1}+1) \rfloor + g - Qa_i$. If $\Delta V_i \geq Q$, then the solution doesn't exist. Stop.
   4.5. If $\Delta V_i \leq z$, then it is needed to minimally increase ΔV_i. $\Delta V_i := \lfloor Qa_i + z \rfloor + g - Qa_i$. If $\Delta V_i \geq Q$, then the solution doesn't exist. Stop.
   4.6. If $i < l$, then go to Step 4.4.

5. Based on (4), ensuring $\Delta M = l$ by maximal possible increasing, if needed, the $\Delta V_i > 0, i = \overline{1, l}$ values.
   5.1. $\Delta V := \Delta V + \Delta V_1 + \Delta V_2 + \Delta V_3 + \ldots + \Delta V_l$. If $\Delta V > \Delta U$, then the solution doesn't exist. Stop.
   5.2. If $\Delta V = \Delta U$, then the solution is obtained. Go to Step 7.
   5.3. $y := \Delta U - \Delta V, i := l$. If $Q - \Delta V_i > y$, then $\Delta V_i := \Delta V_i + y$ and the solution is obtained. Go to Step 7.
   5.4. $h := \Delta V_i, \Delta V_i := \lceil Qa_l + Q \rceil - g - Qa_i, y := y - \Delta V_i + h$. If $l = 1$, then it is needed to increase the values of $\Delta V_i, i = \overline{l+1, n}$.

Go to Step 6.
5.5. $i := i-1; h := \Delta V_i; \Delta V_i := \lceil Qa_i + [\Delta V_{i+1}(a_i+1) - Q(a_i - a_{i+1})]/(a_{i+1}+1)\rceil - g - Qa_i$. If $\Delta V_i < Q$, then:
5.5.1. If $\Delta V_i > h+y$, then $\Delta V_i := h+y$ and the solution is obtained. Go to Step 7.
5.5.2. $y := y - \Delta V_i + h$ and go to Step 5.8.
5.6. If $Q > h+y$, then $\Delta V_i := h+y$ and the solution is obtained. Go to Step 7.
5.7. $\Delta V_i := \lceil Qa_i + Q\rceil - g - Qa_i, y := y - \Delta V_i + h$.
5.8. If $i > 1$, go to Step 5.5.

6. Based on (3), ensuring $\Delta M = l$ by the maximal possible increase of the $\Delta V_i \geq 0, i = \overline{l+1, n}$ values.
6.1. $x := min\{\Delta V_i, i = \overline{1,l}\}, i := n, h := \Delta V_i$. If $x > h+y$, then $\Delta V_i := h+y$ and the solution is obtained. Go to Step 7.
6.2. $\Delta V_i := \lceil Qa_i + x\rceil - g - Qa_i, y := y - \Delta V_i + h$.
6.3. If $i = l+1$, then the solution doesn't exist. Stop.
6.4. $i := i-1, h := \Delta V_i; \Delta Vi := min\{\lceil Qa_i + x\rceil; \lceil Qa_i + \Delta V_{i+1}a_i/a_{i+1}\rceil\} - g - Qa_i$. If $\Delta V_i > h+y$, then $\Delta V_i := h+y$ and the solution is obtained. Go to Step 7.
6.5. $y := y - \Delta V_i + h$. Go to Step 6.3.

7. Determining the $V_i, i = \overline{1,n}$ values. $V_i := Qa_i + \Delta V_i, i = \overline{1,n}$. Stop.

The obtained values of $V_i, i = \overline{1,n}$ can be checked by applying the Hamilton method. It should be noted that the affirmations "the solution doesn't exist" in the A_{HL} algorithm are approximate, but very close to reality for $g = 1$. Parameter g is an integer, the value of which influences the minimal difference among the $x_i/V_i - x_{i+1}/V_{i+1}, i = \overline{1, n-1}$ ones: the larger the value of g, the larger the mentioned difference. At the same time, the smaller the value of g, the higher the probability that the solution will be obtained.

Algorithm A_{HL} was implemented in the computer application SIMAP. Examples 1 and 2 using SIMAP are described below.

Example 1 regarding the generation of a Hamilton apportionment which fully favors large beneficiaries. Initial data: $M = 279; n = 20; \Delta M = 10; V = 20000; g = 1$; the $x_i, i = \overline{1,n}$ values specified in Table 1.

Some results of calculus using SIMAP are systemized in Table 1.

Table 1. Calculations for the apportionment to Example 1

i	V_i	$10^{-7}x_i/V_i$	i	V_i	$10^{-7}x_i/V_i$
1	2145	139860	11	932	139485
2	1931	139824	12	789	139417
3	1788	139821	13	718	139276
4	1645	139818	14	575	139130
5	1574	139771	15	504	138889
6	1431	139762	16	433	138568
7	1360	139706	17	289	138408
8	1289	139643	18	217	138249
9	1146	139616	19	145	137931
10	1003	139581	20	86	116279

Example 2 regarding the generation of a Hamilton apportionment which fully favors large beneficiaries. Initial data are the same as in Example 1 with the only difference that $g = 3$. Some results of calculations using SIMAP are systemized in Table 2.

Data of Tables 1-2 were checked – the obtained apportionments are Hamilton ones. At the same time they comply with requirements (1). Thus, they fully favor large beneficiaries.

Comparing data in Tables 1 and 2, one can see that the obtained values of V_i and $x_i/V_i, i = \overline{1,n}$ differ. Using different values of g, one can obtain different solutions.

The minimal difference among the $x_i/V_i - x_{i+1}/V_{i+1}, i = \overline{1,n-1}$ ones is equal: to 3 if $g = 1$ and to 172 if $g = 3$. So, it is confirmed the fact that the larger the value of g, the larger the mentioned difference. Thus, if it is needed to increase this difference, one has to increase the

Table 2. Calculations for the apportionment to Example 2

i	V_i	$10^{-7}x_i/V_i$	i	V_i	$10^{-7}x_i/V_i$
1	2113	141978	11	932	139485
2	1904	141807	12	791	139065
3	1765	141643	13	722	138504
4	1626	141451	14	580	137931
5	1558	141207	15	512	136719
6	1419	140944	16	462	129870
7	1350	140741	17	318	125786
8	1281	140515	18	247	121458
9	1141	140228	19	175	114286
10	1001	139860	20	103	97087

value of g. But the value of g is limited from above by the value of $\lceil Q/n \rceil$ (approximately). In Examples 1 and 2, one has $Q = V/M = 20000/279 \approx 71.7$ and $\lceil Q/n \rceil = \lceil 71.7/20 \rceil = 4$. However, the attempt to obtain the solution using SIMAP for initial data of Examples 1 and 2 at $g = 4$ was unsuccessful.

5 Conclusions

In order to determine Hamilton apportionments which fully favor large beneficiaries, the A_{HL} algorithm was elaborated. It guarantees the solution (if it exists), regardless of the value of n. Two examples of generating of such apportionments at $n = 20$ using the computer application SIMAP are described. In this context, it should be noted that in all 25 million variants of initial data with $n = 20$, for which the $V_i, i = \overline{1,n}$ values were generated stochastically at uniform distribution, none of the Hamilton apportionments fully favors the beneficiaries.

At the same time, it was identified that the results of calculations considerably depend not only on initial data $V, n, 1 \leq \Delta M \leq n-1$ and

$x_i, i = \overline{1,n}$, but also on parameter g value of the A_{HL} algorithm. The higher the g value ($1 \leq g \leq \lceil Q/n \rceil$), the larger the minimal difference among the $|x_i/V_i - x_{i+1}/V_{i+1}|, i = \overline{1,n-1}$ ones. But the maximal value of g, for which it is possible to obtain the solution, strongly depends on the value of ΔM, being small at small or large values of ΔM and large – at medium values of ΔM in the interval $[1; n - 1]$.

References

[1] M.L. Balinski, H.P. Young. *Fair Representation: Meeting the Ideal of One Man, One Vote.* 2nd ed. Washington, DC: Brookings Institution Press, 2001.

[2] U. Kohler and J. Zeh. *Apportionment methods.* The Stata Journal, 2012, 12(3), pp. 375–392.

[3] M. Gallagher. *Proportionality, Disproportionality and Electoral Systems.* Electoral Studies, 1991, 10(1), pp. 33-51.

[4] E. Huntington. *A new method of apportionment of representatives.* Quart. Publ. Amer. Stat. Assoc., 1921, 17, pp. 859-1970.

[5] I. Bolun. *A criterion for estimating the favoring of beneficiaries in apportionments.* Proceedings of Workshop on Intelligent Information Systems WIIS2020, December 04-05, 2020. Chisinau: IMI, 2020, pp. 33-41.

[6] I. Bolun. *Total favoring in proportional apportionments.* Journal of Engineering Science, XXVIII, no. 1, 2021, pp. 47-60.

Ion Bolun

Technical University of Moldova
E–mail: ion.bolun@isa.utm.md

On Classification of 17th Century Fonts using Neural Networks

Tudor Bumbu

Abstract

This paper presents a solution on how to classify the fonts in the 17th century Romanian Cyrillic documents. This solution is based on training an artificial neural network model.

Keywords: old books, OCR, font classification, neural networks, Romanian Cyrillic.

1. Introduction

Some Romanian Cyrillic documents of the 17th century require particular models at optical character recognition (OCR) because some printing houses used different *character printing styles* (fonts) than others [1].

The problem of identifying and classifying the font in a document printed in the 17th century can be formulated as follows: *Given a document X from the 17th century printed in Cyrillic Romanian and a set N of OCR models trained on documents of the 17th and 18th century, choose the most appropriate model from N for X.*

A trivial solution would be to recognize a sample (a page snippet) from document *X* using all models in *N* and, based on the results, to choose the model that gives the highest accuracy (the best result). This solution would be easy to implement, but the time complexity is too big, as we have to load each model separately. Model upload time and sample recognition can exceed 2 minutes depending on page size and if we have had 10 different models, we would have to wait for approx. 20 minutes to find the right model each time we want to recognize a document of this kind.

The proposed solution will be to train a neural network with samples from several Romanian documents printed in the 17th century at different

printing houses. A neural network will learn from a training dataset consisting of tuples of *image characters* and its *class* (*0*, *1*) in order to be able to further classify a new sample.

In the next section, we will describe the selected document samples aiming at creating the dataset.

2. Dataset resources

The Romanian Cyrillic alphabet was used at printing houses in regions as *Iași, Bucharest, Târgoviște, Belgrade (Alba Iulia), Uniev (Cernăuți), Sas Sebeș, Snagov, Buzău.* Each of these regions had at least one printing house in the 17th century.

The data set is created from 10 scanned books, selected from the digital library of Romania (http://digitool.bibnat.ro). In the selected books, two distinct sets of characters were observed. Therefore, the books were divided into two classes depending on their font style: books no. 1, 8, 9, 10 were put in set *A* and the books with no. 2, 3, 4, 5, 6, 7 were put in set B. When forming the dataset, 13 pages were included in *A*, and 9 pages in *B*. Figure 1 shows two samples from each set *A* and *B*. The two main letters which differ in both samples are *m* and *з (t and z in Lattin).*

Figure 1. A sample from *A* (on the left) and one sample from *B.*

In the next section, let's take a quick look at the main tasks in the process of creating the dataset: segmentation of blocks of text from the selected pages; detection of the individual characters in text blocks; clustering the characters and forming the training and testing data sets from the detected characters.

3. Creating the dataset

Below there are described tools for segmenting the regions of text (text blocks) in the selected pages, methods for identifying the individual characters in a text block, and approaches to group characters in similar groups in order to create a better dataset.

3.1 Segmentation of text blocks

From the pages prepared for the dataset, we will start segmenting and cutting fragments of text (text blocks) using *Detectron2* [2] segmentation tool trained on the *PrimaLayout* dataset [3].

In the *PrimaLayout* model, the text portions are labeled with the label "*TextRegion*". We will segment and cut more than one block of text from a single page. For this reason, two blocks of text may contain the same characters, and similar examples of training and testing may appear in our dataset.

After segmentation, the text blocks are to be cut and placed in a list of text block snippets. On average, 4 fragments with blocks of text were obtained for each of the 22 selected pages. These fragments are images. In the next subsection, we will identify and cut the characters from text block snippets. The character set consists of *letters*, *punctuation marks*, *accents*, *some lines in the outline of tables*, *stain pixels*, or *page noise*.

3.2 Detecting individual characters of text blocks

What we need to make sure is that each image, independent of its source, is processed in such a way that the algorithm used to detect the letters can find as many letters as possible. We will convert all images to black and white. As a result, the processed image will consist only of black and white pixels. We can then further optimize the image for letter detection using the *findCountours()* method within *openCV* library (https://docs.opencv.org/4.5.3/). We can then map the contour delimitation boxes to the original image to see what was actually detected. The result of this processing step is shown in Figure 2.

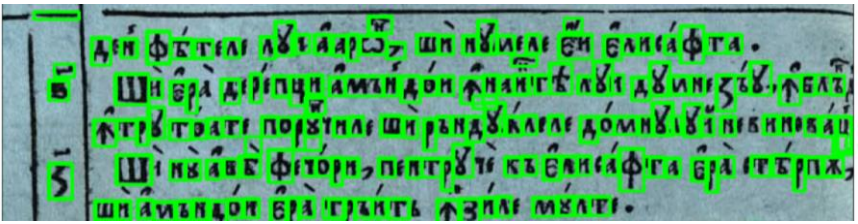

Figure 2. Drawing the contours of the characters (green rectangles) in a block of text from set *A*.

The characters inside green boxes (Figure 2) were cut and saved in two folders, for each font separately. From the first 10 pages of set *A* we obtained 17155 characters, and from *9* pages of set *B* – 8799 characters. Among the extracted characters, some could be elements of noise in the image – spots, accents without the basic character (letter), tilde, etc.

In the next subsection, we will try to remove unnecessary characters and keep only the letters. For this task, we use *K-Means* clustering model implemented in Scikit-learn library (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html).

3.3 Clustering the characters

A set of characters extracted from fragments of text blocks is fine, but it would be good to organize them in such a way that these letters become useful to train a neural network. Some action would be to remove images that do not contain letters; to cluster all the remaining images by clustering (this means that all the letters "ѣ" will form a cluster, all the letters "ъ" form another cluster, and so on).

Before we start clustering, we need to make sure that all the images are the same size. Each extracted character has the size of its boundary box, which varies widely. So, we will resize all the images to the size of *50*50 pixels*.

We will apply *K-Means* clustering because it is simple and fast. The only thing we need to provide is the number of clusters we are waiting for. If *K-Means* would give us a perfect result, then we would end up with the exact number of letters in the Romanian Cyrillic alphabet – 47 letters, but of course, we should also consider some variations on the differences between uppercase and lowercase. Each of them can appear as a

lowercase or uppercase letter, which means that, in total, we would expect 94 clusters. There will also be punctuation marks and noises in the data that have been detected. Therefore, we will set 100 clusters. This is more than we need, but it will be easier to merge clusters later than to separate them.

After analyzing the clusters and deleting characters that are not *letters* and *punctuation marks*, we will place them in two bigger clusters, one for set *A* and another cluster of characters for set *B*. The result of this process is two folders called *fontA* and *fontB*. There are more than 10,000 characters in the *fontA* (set *A*), and the *fontB* (set *B*) folder contains 8,775 characters.

To form a well-organized dataset, we will convert the dataset to *IDX* format [X]. Since we want to train a neural network, we should divide the images into a set of training and testing data. For this, we will move *1/3* of characters from set *A* and *B* to a test folder. The output will be 4 folders: 2 folders for image characters – one for training and one for testing, and another 2 folders with the labels corresponding to the character folders. The labels with the character class will be the values *0* and *1*, where *1* is the class for a character from set *A* and *0* is the class label for a character from *B*. After counting the examples in the dataset, we have 21,212 *training examples* and 9,093 *test examples*.

In the next section, we will train a multilayer neural network with the data set prepared at this stage.

4. Train the neural network

We will train a multilayer neural network (NN) to classify the characters in two different fonts. The architecture of the NN model will be characteristic of binary classification and will be implemented using *Tensorflow* (https://www.tensorflow.org/guide/keras/sequential_model).

First, we need to upload the dataset we previously saved in the *IDX* data format. In Figure 3 we can see some examples in the training dataset.

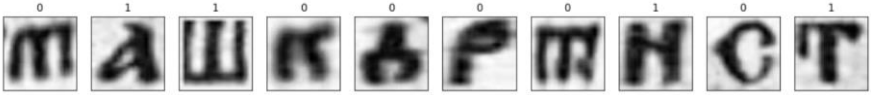

Figure 3. Examples taken from the training dataset, where each image character has its class label – *0* or *1*.

It's time to build the neural network. We will start with a transformation of the input data from an *x* with *y* matrix into a vector of length *x* * *y*. Thus, we have an input layer of 2,500 neurons. Then we will add a hidden layer with 128 neurons with the *ReLU* activation function (which is completely connected to the last layer. The last layer (output layer) contains a single neuron and a sigmoid function for its activation. At the training phase, we set 300 epochs, and the training lasted about 55 minutes without GPU.

The training went well and we obtained an accuracy of 96.7%. Based on the confusion matrix (Figure 4), we calculate the classification error to be 3.3%.

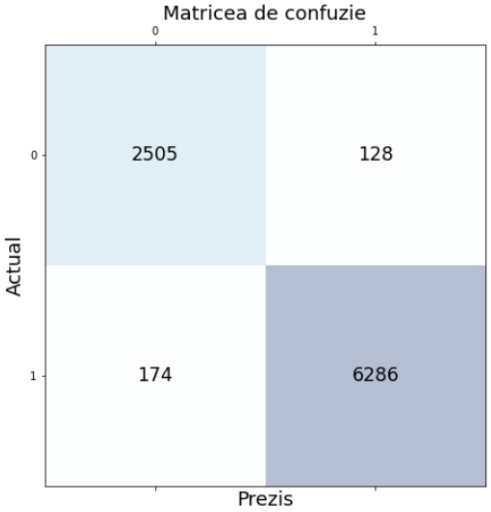

Figure 4. Confusion matrix based on the test data set.

5. Conclusion

In this paper, we have trained a neural network to classify two different fonts used at printing Romanian Cyrillic documents in the 17th century. We presented the main steps in preparing the dataset for this training and the resources we used. The obtained NN model can be used to identify the best OCR model to use on a particular document depending on its font style.

References

[1] T. Bumbu, S. Cojocaru, A. Colesnicov, L. Malahov, and Ș. Ungur. *User Interface to Access Old Romanian Documents*. In: Proceedings of the 4th Conference of Mathematical Society of Moldova CMSM4'2017, June 25-July 2, 2017, pp. 479-482.

[2] R. Girshick, I. Radosavovic, G. Gkioxari, P. Doll´ar, and K. He. *Detectron*. https://github.com/facebookresearch/detectron, 2018.

[3] C. Clausner, A. Antonacopoulos, and S. Pletschacher. *Prima Layout* ICDAR2017 competition on recognition of documents with complex layouts-RDCL2017. In: Proceedings of 14th International Conference on Document Analysis and Recognition (ICDAR), volume 1, pages 1404–1410, 2017.

Tudor Bumbu [1, 2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E-mail: tudor.bumbu@math.md

[2]State University of Tiraspol
E-mail: bumbutudor10@gmail.com

Discover the Moldovan Cultural Heritage through e-Moldova Portal by Using Crowdsourcing Concept

Olesea Caftanatov, Tudor Bumbu, Lucia Erhan, Iulian Cernei, Veronica Iamandi, Vasile Lupan, Daniela Caganovschi, Mihail Curmei

Abstract

Our research aims to analyze the needs of the people and, by using modern technologies, to develop a portal that will use the wisdom of crowds to contribute to the development of the Republic of Moldova. In this paper, we present the first version of e-Moldova portal.

Keywords: crowdsourcing, portlets, digital cultural heritage, Flarum core.

1. Background research

Technology is present more than ever in our lives and it is difficult to imagine our daily routine without it. Over the years, technology has paved the way for multi-functional devices. As a result, it made our lives easier, faster, better, and more fun. The Republic of Moldova still has work to do in terms of implementing modern technologies in everyday life, but it has a great potential to manifest itself in the IT field in the near future. Our team pursues the following objectives:

- researching the needs of people and bringing new technologies to make our country a prosperous one in the near future;
- searching fields in which our expertise will be more useful for our country;

- collecting, creating, and managing informational resources regarding our country, also providing free access to them;
- digitization of our cultural heritage by involving crowds wisdom in our project;
- developing tools that will help the mass to rediscover our country.

In this paper, we will present the results of one year of our work in pursuit of these objectives and describing the first version of our project.

2. E-Moldova Portal

Our team also brings its contribution to the promotion and development of the Republic of Moldova through the development of the e-Moldova portal [1], a platform dedicated to digitization of our cultural heritage and complementing it with new knowledge by involving the wisdom of crowds in exploring "portlets" on history, culture, geography, demography, politics, economics, career, education, wellness, press, informational technologies, and, of course, the digital thesaurus (see Fig.1).

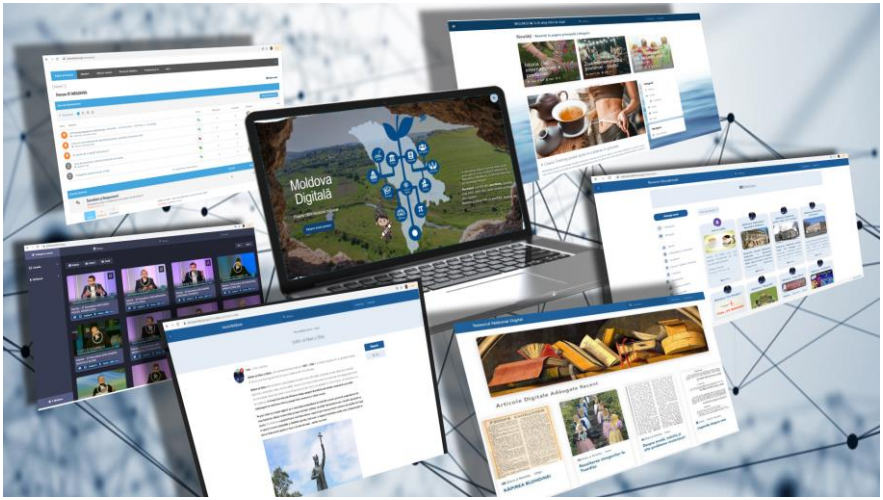

Figure 1. E-Moldova portal main interface with six portlets

According to the Lexico dictionary powered by OXFORD [2], the term *"portlet"* in the informational field means *"an application used by a portal website to receive requests from clients and return information"*.

Our team essentially sees a portlet as a specialized content area within a web page that occupies a small place and serves as a hotspot leading towards a large body of information (the word ***portlet*** has the same meaning as the word ***door***). The nature of information can variate from typical posts, video lessons, galleries, and access to other resources or even tools. Nevertheless, all portlets are designed to use the wisdom of crowds in order to further evolve. In other words, our project is based on the concept of "*crowdsourcing*" developed by James Surowiecki [3].

The meaning of this concept can be conveyed by the formula: if two heads are better than one, a hundred heads will be great. We constantly encounter applications based on this approach. An eloquent example of the use of mass intelligence is Wikipedia.

Thousands of Wikipedia users have created an encyclopedia that, according to [4, 5], is considered to be a good source of almost the same accurate information as Britannica. One of our objectives is to create a tool that will enable crowds' intelligence to create a body of knowledge about the Republic of Moldova presented in a digital format.

To this day, we have initialized 12 portlets that can be accessed through the e-Moldova portal (see Fig. 2). The portal's interface was intended to be as simple as possible: a menu is displayed as a grape, and each berry represents one of the portlets.

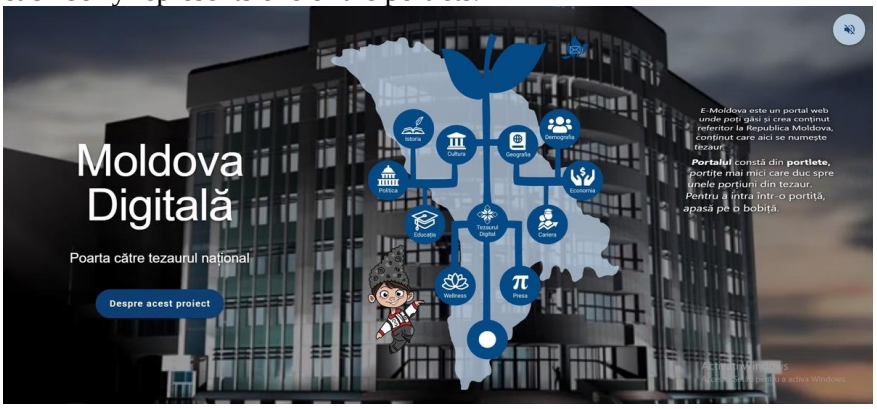

Figure 2. The interface of the e-Moldova portal

Each berry has a toolkit with a short description of portlets. In the background, we play a video with places that can be seen in our country or events that took place in it. Thus, the background is a component of the user interface that adds to the knowledge about Moldova.

The portal's layout consists of 3 sections. On the left side, we have the title and a button that redirects to a discussion board, where anyone can leave some feedback for improvement of the portal or even its portlets. On the right side of the layout, we displayed an animated text with some tips on how to access our portlets. In the upper right corner, we added a voice button that recites the text in 3 languages (Ro, Ru, and Eng). The main part of the layout is the center, where our menu is displayed as a grape on the outlined map of our country. In addition, near the right leaf, we have a pigeon with the envelope, for the case when the accessed users wish to leave a message to the administrator.

Another digital element is in the process of developing, it is about the Guguța avatar, and this animated character will navigate through our interface and help users to better understand our project.

3. Tools for Portlet development

To develop our portlets, we used Flarum discussion platform [6] as a core and Wordpress content management system [7]. Flarum is built with Laravel framework [8] so it's quick and easy to deploy. Moreover, its interfaces are powered by Mithril, a performant JavaScript framework for single-page applications. Flarum's architecture is flexible, and it is equipped with a powerful Extension API. We easily can extend, customize or even integrate our extensions to the base platform. All portlets, except IT forum portlets, were developed using Flarum Core. Regarding the IT forum portlet, we used the WordPress platform. We also intend to use a wiki engine [9] to bring the possibility for users to collaboratively edit the content (the articles).

4. Portlets

As we have mentioned before, we have initialized 12 portlets that can be accessed through the e-Moldova portal, but in this paper, we described only five of our portlets: *IT Forum, Education, Press, Digital Thesaurus, and Wellness*. We chose to describe these five portlets because they have

more advanced interfaces and functionalities. The other initialized portlets have the basic functionality for content creation and management. Going forward, we intend to add specific functionality for each one of them.

4.1 IT Forum Portlet

One of the first portlets initialized was the IT forum. This portlet brings together a wide range of current information technologies, including useful information for pupils, students who are interested in learning a programming language or learning more about some branches of IT (see Fig. 3).

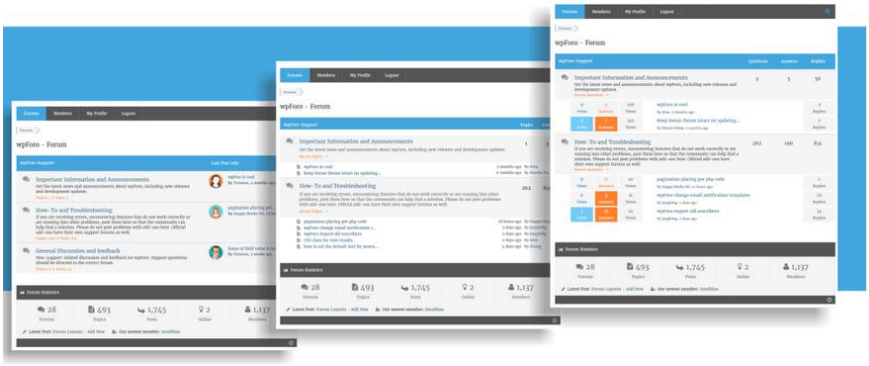

Figure 3. IT Forum portlet homepage interfaces

As a core for IT Forum, we use wpForo [10] plugin from WordPress platform made by gVectors Team. wpForo has become a new generation of Forum Software. It has multi-layouts such us: The “Extended”, “Simplified” and “Question & Answer”. Layouts fit almost all types of discussions needs.

Beautiful, modern, and informative profile system, with member pages as a statistic, bio, settings, activity, and subscriptions. It has a user rating system based on several posts. Nice Badges and Member Rating Titles per reputation level. Powerful moderation tool, fully customizable.

Our team for one year created 580 topics with 665 posts organized in 47 under IT forums. Each post that was created was checked on plagiarism by using PREPOSTSEO Tool [11]. Only content that has an average of 70% unique was approved and published. Thus, through passion, we develop a community for IT geeks.

4.2 Education Portlet

The Education portlet (see Fig. 4) is a collection of information resources, intended for four categories of users, grouped in distinct communities such as the community of teachers, parents, pupils, and the community of students. Resources are selected for user education or as an aid in choosing a school or university; courses, sports clubs; the choice of useful tools in education, etc.

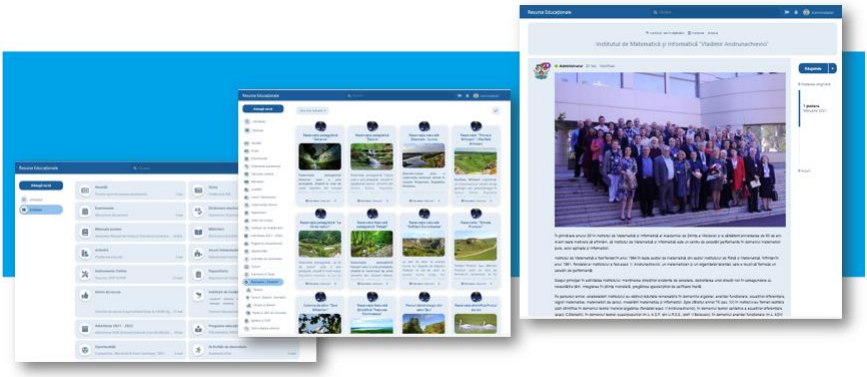

Figure 4. Educational portlet interfaces

Now, the Educational portlet has over 700 information resources presented in the form of cards, which continue to grow every day. Guests have access to sources and even the ability to post a source that may be useful to other users, or may leave feedback in the form of comments. For the Educational portlet, we allow any type of added resources, the minimal requirements are:

- useful resource;
- no uncensured content is allowed.

4.3 Press Portlet

This portlet's mission is to increase the impact of the independent press in the Republic of Moldova and to contribute to the creation and consolidation of the open society (see Fig. 5).

We believe that anyone who has the ability can create an article even if he/she is not a journalist by profession. At some point, every person can contribute with any kind of information.

Our team can moderate the posts created by the masses and generate from the top rating posts a digital newspaper. However, because of little human resources, we started with a compilation of the best TV Shows on three popular Channels: “Publika TV”, “TVC21”, and “10TV”. All the collected resources can be analyzed and get a place in the rating system. In such a way the more useful TV Shows will be posted on top of the list.

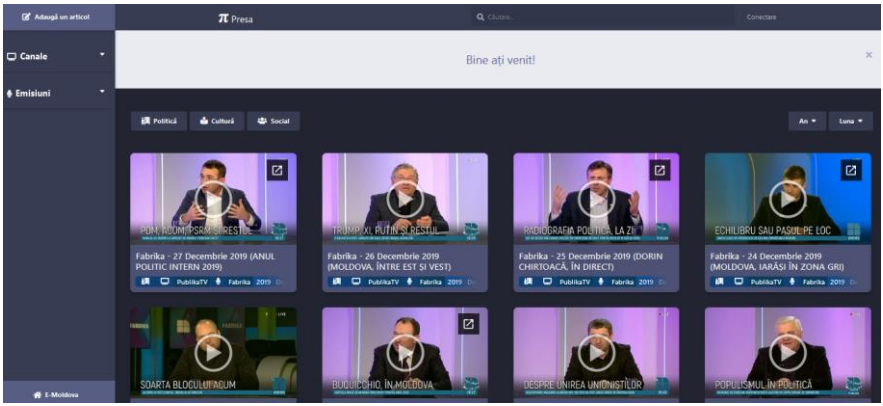

Figure 5. Press portlet homepage interface

4.4 Digital Thesaurus Portlet

The Digital National Thesaurus portlet aims to digitize articles that are part of the Moldovan cultural heritage, using sophisticated information technologies to help you study and learn about the country's past and ancestors (see Fig. 6).

We aim to provide access to articles from the following collections: newspapers and magazines, archival documents, books and manuscripts, collections gathered and selected from past centuries. The word "digitization" refers to a technological process, where an item from some collections is converted and transformed into a digital item, which can then be placed on the Internet.

Such an article can be one or more pages from a newspaper, a magazine or a book, a historical document, but maybe even a photo from a Moldovan village, and so on.

The portal will ensure access to newspapers and magazines printed in Moldovan Cyrillic alphabet. We consider newspapers and magazines as a

priority collection because they are about the lives of Moldovans with "many truths left in the past".

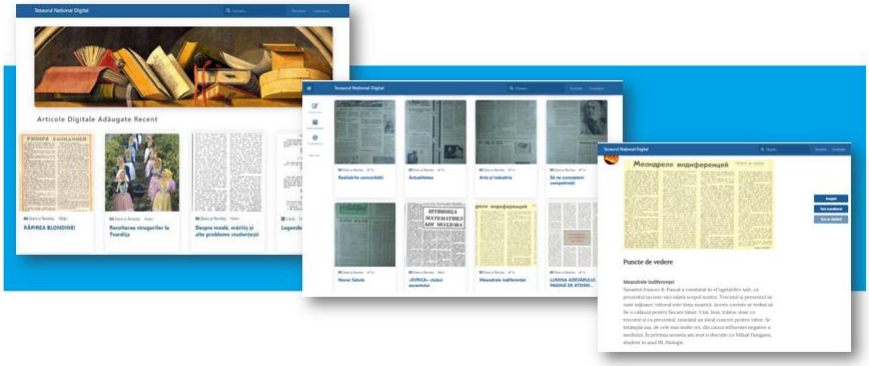

Figure 6. Digital Thesaurus portlet interfaces

Newspapers and magazines from the 20th century are kept in many libraries in the country, and among them, there are the following: the National Library of Moldova, the National Archive of the Republic of Moldova, the Central Scientific Library "Andrei Lupan", the USM Library, etc. The result of this project is the national digital hoard consisting of digital articles.

This portlet differs from the other portlets because it includes technology for digitization. The technology consists of a tool pack for image preprocessing, layout analysis and segmentation, optical character recognition, and transliteration from the Moldovan Cyrillic alphabet to Latin.

4.5 Wellness Portlet

The Wellness portlet was created to inspire people to choose a healthier lifestyle. Through this portlet, our team distributes some results from studies done on various dimensions of well-being what we call Wellness (see Fig. 7).

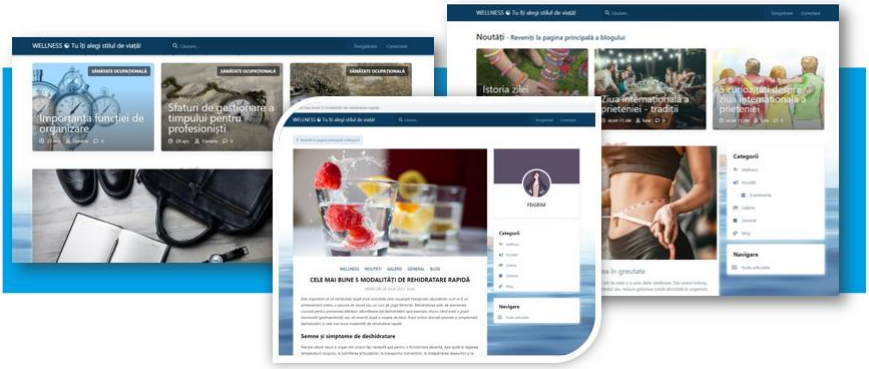

Figure 7. Wellness portlet interfaces

Because wellness is more than getting rid of an illness, wellness is a permanent dynamic process, to become aware, responsible, and to make decisions that contribute to well-being. It refers to change and growth, through which the physical, intellectual, emotional, social, spiritual, occupational levels, as well as the level of the environment, develop. Every dimension is equally vital in the pursuit of optimal health. It is the perfect balance between mind, body, and spirit.

In total, we have created over 230 topics for discussion that can be commented on, appreciated, and distributed on various social networks. The portlet provides free access to nutritional information, recipes, and more. All posts were checked by a plagiarism checker, those that have plagiarized content more than 30%, we return to the author and ask them to redo. Additionally, if there are some scientific articles in other language but are very interesting, we translate them and post. Moreover, at the end of the post, we write that it was a translated work and we share the link to the original article.

5. Conclusion

In this paper, we presented the e-Moldova project that uses the crowdsourcing concept and modern technologies to develop a portal with its portlets. Each portlet leads to a specific field about the Republic of Moldova, be it regarding the economical field or political, etc. In this article, we described the purpose of the main portal and shared some ideas

about the first five portlets. Currently, the content is being created and managed by our team. It seems to be a normal thing in the initial phase. However, we intend to delegate this work to the users and in this manner implement the crowdsourcing approach.

Acknowledgments. We would like to express our sincere gratitude to Dr. Ioachim Drugus, the owner and Dev Lead of e-Moldova portal © (also referenced as e-Moldova ©, or Digital Moldova ©), for continuous support and for giving us the opportunity to study, research and participate in the development of the portal. We hope that those who learned about this project and were inspired by our activity will try to get involved in any form – through feedback, by sending an article for publication, or by suggesting a useful resource. On this occasion, we would like to also express our gratitude to the people on the project "Intelligent information systems for solving ill-structured problems, processing knowledge and big data", who support us, the young researchers, engineers, inventors, that still cannot produce something to be sold.

References

[1] *E-Moldova Portal*, official website: https://emoldova.org/.

[2] *Lexico Dictionary*, website link: https://www.lexico.com/definition/portlet.

[3] J.Surowiecki. *The Wisdom of Crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* Doubleday, 2004, pp. 296.

[4] D. Terdiman. *Study: Wikipedia as accurate as Britannica.* Digital post on Cnet website. December 16, 2005. Accessed on August 1, 2021. https://www.cnet.com/news/study-wikipedia-as-accurate-as-britannica/.

[5] N. Wolchover. *How accurate is Wikipedia. Digital post on Live Science Website*. January 24, 2011. Accessed on July 15, 2021. https://www.livescience.com/32950-how-accurate-is-wikipedia.html.

[6] *Flarum*, official website. https://flarum.org/.

[7] *WordPress*, official website: https://wordpress.com/.

[8] *Laravel*, official website: https://laravel.com/docs/8.x.

[9] *Wiki Software*. Wikipedia, https://en.wikipedia.org/wiki/Wiki_software.

[10] *wpForo Plugin on WordPress*, official website: https://wpforo.com/.

[11] *Plagiarism checker Tool*, official website: https://www.prepostseo.com/plagiarism-checker.

Olesea Caftanatov[1], Tudor Bumbu[2], Lucia Erhan[3], Iulian Cernei[4], Veronica Iamandi[5], Vasile Lupan[6], Daniela Caganovschi[7], Mihail Curmei[8]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science; SRL EST Computer
E-mail: olesea.caftanatov@math.md

[2]Vladimir Andrunachievici Institute of Mathematics and Computer Science; SRL EST Computer
E-mail: tudor.bumbu@math.md

[3]EST Computer SRL
E-mail: ladyblackserinity@gmail.com

[4]EST Computer SRL
E-mail: iuliancernei@gmail.com

[5]Vladimir Andrunachievici Institute of Mathematics and Computer Science; SRL EST Computer
E-mail: veronica.gisca@gmail.com

[6]EST Computer SRL
E-mail: lupan.01@mail.ru

[7]State University of Moldova; SRL EST Computer
E-mail: cda3721@gmail.com

[8]EST Computer SRL
E-mail: mcurmei@mail.ru

Virtual Reality FantasyShooter

Teodora Cătălina Călărașu, Adrian Iftene

Abstract

Virtual Reality (VR) is a domain that has grown a lot in the previous years and opened new paths in the way that people interact with technology. It creates a simulated environment in which a person can have similar experiences like those in the real world or completely different ones, depending on the purpose of that simulation. Virtual Reality is used in different domains from various types of training, education, medicine to unique ways of relaxation and entertainment. The idea of having a game implemented in Virtual Reality and it being something more complex than an application made for a simple Google Cardboard sounded like an interesting and unique challenge thus the concept for VR FantasyShooter was born. VR FantasyShooter can be used for recreational purposes and can also help in training certain reflexes since it puts a lot of accent on the coordination between the hands and the eye. It also offers the possibility of walking around that world without the controllers but the movement via the joysticks was implemented in order to not depend on the limitation of the physical space around you.

Keywords: Virtual reality; Shooter game; 3D modeling.

1. Introduction

Virtual Reality (VR) is a domain that has grown a lot in the previous years and opened new paths in the way that people interact with technology. It creates a simulated environment in which a person can have similar experiences like those in the real world or completely different ones, depending on the purpose of that simulation. VR is used in different domains from various types of training, education, medicine to unique ways of relaxation and entertainment [1], [2]. What's the most interesting part about this area is the way that it developed from a simple concept of

finding a technique which could give a static image a 3D depth by using a stereoscope [3] to a simple cardboard that was capable of making people explore different areas by simply turning their head and even further to devices that could track your whole body movements and offer a complete experience using the previously mentioned techniques. The different ways of implementing this virtual reality in order to create an immersive experience for a user is a topic that has interested us since the apparition of the first modern VR headsets and led to a growing curiosity and eagerness to learn more about it.

The idea of having a game implemented in VR and it being something more complex than an application made for a simple Google Cardboard sounded like an interesting and unique challenge thus the concept for VR FantasyShooter was born. VR FantasyShooter can be used for recreational purposes and can also help in training certain reflexes since it puts a lot of accent on the coordination between the hands and the eye. It also offers the possibility of walking around that world without the controllers but the movement via the joysticks was implemented in order to not depend on the limitation of the physical space around you.

2. Similar Applications

At this moment the developed application is somewhat unique on the VR market since the accent was on integrating different PC games mechanics inspired by MOBAs[1] and PVEs[2] which usually aren't present together, in such a manner to create an easy transition between the 2 devices. However, there are certain applications that could be considered similar if we are only taking into consideration certain movements or actions that the user is allowed to do.

2.1 Beat Saber

Beat Saber[3] is a VR exclusive game that was designed around the idea of making exercises more pleasurable for people, therefore it is considered to be a fitness application [4]. The premise of the game is simple but yet

[1] Abbreviation for Multiplayer online battle arena
[2] Abbreviation for Player versus environment
[3] Beat Saber: https://beatsaber.com/

captivating, staying on a platform with two different color lightsabers and slashing the cubes in the direction indicated. Even if it doesn't require room scaling since you only have to stay in a small area, it encourages different types of body movements and puts a lot of emphasis on hand-eye coordination. The asymmetric way in which cubes come and as well the different directions that they have to be slashed represents a good way of practicing your concentration and your reflexes.

Figure 1. Beat Saber Gameplay[4].

2.2 Superhot

Superhot[5] is an application that was initially developed for PC and later adapted for the VR headsets [4]. The basic idea of the game sounds plain, a shooter, but the way that it is built makes it unique. Each level is constructed like a puzzle where you have to shoot different targets, but the way you do it is the important part. When you are not moving, the time stops giving you time to make a strategic decision in order to complete the stage. The more you move the faster everything is, so certain actions might lead to an imminent loss. Making the right choice in different situations whether that is to just look around, to pick a specific weapon to

[4] https://www.igdb.com/games/beat-saber
[5] https://superhotgame.com/

defend yourself, or hide together with good reflexes and an analytic eye are mandatory in this time-manipulation type of application.

The level design is minimalist in order to move the focus to what's really important but easily manages to immerse the player in this new environment. Combining this with the tactical decisions you have to take with the way that the game mechanics are built and with the characteristic sounds, this application manages to build a complex puzzle game delivered in the form of a fast-paced shooter.

Figure 2. Superhot Gameplay[6].

3. Proposed Solution

VR FantasyShooter being developed for a new type of device which is the Oculus Quest, the implementation was accompanied by research on how this particular headset works. One interesting finding is the way that the VR devices read the canvases from Unity which will later be discussed. The most relevant parts were the performance of it that could be affected by the 3D models since they needed some sort of optimizations and how the controllers work. There are three modes in which we can interpret the input from the controllers and that is the raw method in which each key has a different name, the individual controllers one which doesn't care which hand it belongs to and the one in which we consider the controller

[6] https://www.roadtovr.com/superhot-vr-2-million-rift-vive-psvr-quest/

combined, as a pair. The most used ones in the application are the raw input one and the combined one.

The application is structured in four scenes which gives a logical flow. The first one is the main screen you access when you first enter the game, the second one is the one corresponding to the game itself and the last two represent the winning and the losing state. All of these use an asset for the skybox called BOXOPHOBIC but different settings for it were created in order to create different atmospheres for the player. This asset was downloaded from the Unity store[7].

3.1 Menu Scene

The menu scene is the first one that the player will see when first opening the application (see Figure 3). It has a simple configuration, only containing a canvas with text, buttons, images, and also an Oculus Quest player model without a rigid body since no collision with it is needed for this. Here the user can choose from two options which are to start the game or exit it.

Figure 3. The start screen.

If they decided to start, then the panel corresponding to this is hidden and a new one will be shown. In this new one, there is the option to either go back or to choose between the three images which will decide the

[7] https://assetstore.unity.com/packages/vfx/shaders/free-skybox-extended-shader-107400

character that the user will use (see Figure 4). After choosing the model, a new scene will be loaded.

Figure 4. The character selection screen.

3.2 Game scene

The game scene is based on three managers:

- Game manager checks the alive player counts and when it reaches 0 the game is lost and it loads the LoseMenu scene. It also verifies if the boss's health has reached 0, if this is the case the game is won and the win screen is loaded.
- Waves manager is responsible for the way that the monster will be spawned around the map. It has defined two classes for waves, each wave having two different spawn points for the monsters. The spawn has a name, a location which is determined by the position of an empty GameObject attached to the script, the types of monsters that can appear from there since there are multiple types, a delay and a chance for spawning a monster which is harder to beat than the normal one. The second class is for the waves of monster management and it contains a list of spawns that can be used in the duration of a wave, a time until the wave is over, and a delay.
- Level manager determines when the boss fight will start. After the wave manager stops spawning monsters and all the monsters are killed the fountain will be destroyed by the script and a loading screen will appear. After a few seconds, the boss model appears in the place of the fountain removed beforehand.

The scene itself has an OVRPlayerController taken from the Oculus asset for the integration part which will be controlled via the controllers and the headset. In addition to the base scripts and the components it had, additional ones were needed. It has a rigidbody that has the Use Gravity checkbox selected, the playerSpecs, and all the character scripts attached. After selecting the character in the menu, the corresponding script is activated in order to have the specific types of spells.

In the scene, there are also present at first all three types of computer players but one of them will be deactivated in the previously mentioned manner. There exist two primary sources of light which combined with the renderer settings will create a foggy somewhat mystical environment. To accentuate it even more there are also different forms of colored particles in the air and shaders applied to different objects like the crystal in the middle which will make it appear and disappear (see Figure 5).

Additional colliders had to be added to the outer walls since the convex ones were preventing the other movable object from changing their position and their rotation as expected.

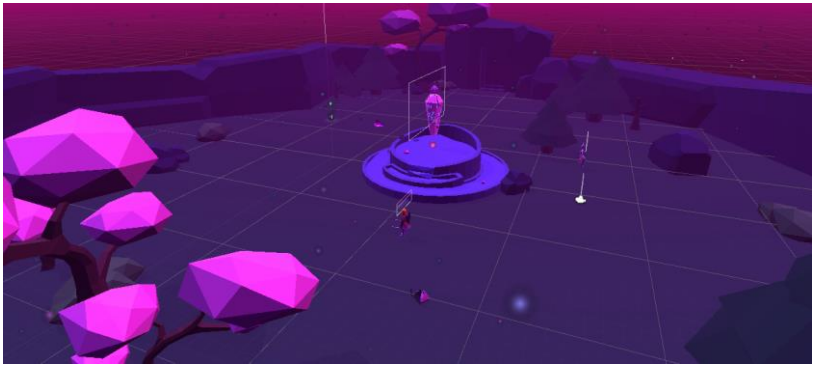

Figure 5. The scene after applying a skybox.

3.3 Win and Lose scene

The last scene which is shown to the user is decided in the game manager script as stated before. Those two are similar to the menu, one only having a canvas with a text and a button that helps the user to go back to the main screen.

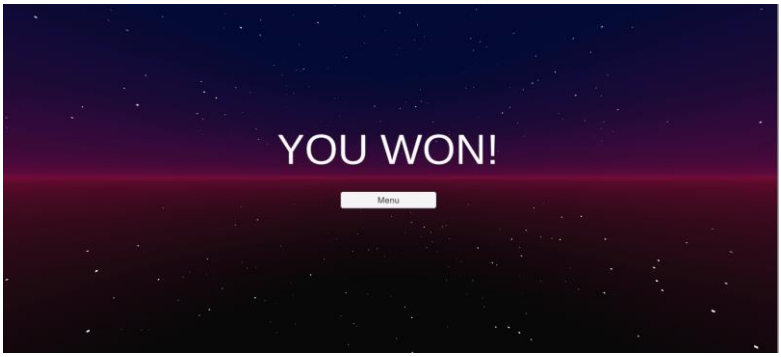

Figure 6. The winning screen.

4. Conclusion

VR is a relatively new domain that developed a lot in the previous years and still continues to expand thanks to the interest that the user manifested in it and into the experiences that it can offer.

VR FantasyShooter is a virtual reality application developed for the standalone headset Oculus Quest which brings the PC games mechanics and possibilities in this new environment. The application represents a good way to relax and entertain yourself and the manner that it is built in makes it intuitive and perfect even for the users that are not used to this kind of new experience. It also brings variety to what a person can play with and this combined with the overall idea of a low-poly arena player versus monster game, makes the application unique on the current Oculus store. The 3D models and the animation which are created by us give a nice personal touch to it and manage to create a mysterious atmosphere in which the user can be easily immersed.

Since it's a new type of technology and many people don't know about it, the resources are limited so there are not many applications currently for Oculus Quest. With VR FantasyShooter we hope to raise the curiosity for VR and increase the interest in developing such concepts.

For the next period, we intend to perform usability tests similar to [5], [6] in order to see user's opinion about this application.

Acknowledgments. This work was support by project REVERT (taRgeted thErapy for adVanced colorEctal canceR paTients), Grant Agreement number: 848098, H2020-SC1-BHC-2018-2020/ H2020-SC1-2019-Two-Stage-RTD.

References

[1] D. E. Gușă, A. Iftene, and D. Gîfu. *Solar System Explorer*. In: 5th Proceedings of the Conference on Mathematical Foundations of Informatics. 3-6 July 2019, Iasi, Romania, 2019, pp. 295-304.

[2] A. Simion, A. Iftene, and D. Gîfu. *An Augmented Reality Piano Learning Tool*. In Proceedings of the 18th International Conference on Human-Computer Interaction RoCHI 2021, 16-17 September, Bucharest, Romania, 2021.

[3] T. Clive. *Stereographs Were the Original Virtual Reality*. Smithsonian Magazine, 2017, https://www.smithsonianmag.com/innovation/sterographs-original-virtual-reality-180964771/.

[4] J. Feltham. *The 25 Best Oculus Quest Games and Experiences*. UploadDVR, 2020, https://uploadvr.com/the-25-best-oculus-rift-games/.

[5] M. Chițaniuc, A. Iftene. *GeoAR-An Augmented Reality Application to Learn Geography*. Romanian Journal of Human-Computer Interaction, vol. 11, issue 2, 2018, pp. 93-108.

[6] M.N. Pinzariu and A. Iftene. *Sphero - Multiplayer Augmented Game (SMAUG)*. In: International Conference on Human-Computer Interaction, 8-9 September 2016, Iasi, Romania, 2016, pp. 46-49.

Teodora Cătălina Călărașu[1], Adrian Iftene[2]

[1]"Alexandru Ioan Cuza" University of Iasi, Romania, Faculty of Computer Science
E-mail: teodora.calarasu@info.uaic.ro

[2]"Alexandru Ioan Cuza" University of Iasi, Romania, Faculty of Computer Science
E-mail: adiftene@info.uaic.ro

Proceedings of Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

Groupoids up to isomorphism of order three with some Bol-Moufang identities

Vladimir Chernov, Valentina Demidova,
Nadeghda Malyutina, Victor Shcherbacov

Abstract

We count the number of non-isomorphic groupoids of order three with some Bol-Moufang identities.

Keywords: groupoid, identity, non-isomorphic, Bol-Moufang identity.

1 Introduction

One of the main questions that mathematics tries to answer is the following: "How much ?"

We count the number of non-isomorphic groupoids of order three with some Bol-Moufang identities.

List of all classical Bol-Moufang identities is given in [1]. We continue the research of groupoids of small order with Bol-Moufang identities [2, 4, 3].

In [4] (see also [5] for quasigroups) it is proved the following theorem in which groupoids with Bol-Moufang identities and groupoids that have equal numerical characteristics are indicated.

From Theorem 1 it follows that groupoids with identity F_1 and identity F_3, groupoids with identity F_7 and identity F_8 have equal numerical characteristics and so on. Notice, identities F_1 and F_3 we name as dual. And so on.

Theorem 1. *For classical Bol-Moufang type identities over groupoids the following equalities are true:*

$(F_1)^* = F_3$, $(F_2)^* = F_4$, $(F_5)^* = F_{10}$, $(F_6)^* = F_6$, $(F_7)^* = F_8$, $(F_9)^* = F_9$, $(F_{11})^* = F_{24}$, $(F_{12})^* = F_{23}$, $(F_{13})^* = F_{22}$, $(F_{14})^* = F_{21}$, $(F_{15})^* = F_{30}$, $(F_{16})^* = F_{29}$, $(F_{17})^* = F_{27}$, $(F_{18})^* = F_{28}$, $(F_{19})^* = F_{26}$, $(F_{20})^* = F_{25}$, $(F_{31})^* = F_{34}$, $(F_{32})^* = F_{33}$, $(F_{35})^* = F_{40}$, $(F_{36})^* = F_{39}$, $(F_{37})^* = F_{37}$, $(F_{38})^* = F_{38}$, $(F_{41})^* = F_{53}$, $(F_{42})^* = F_{54}$, $(F_{43})^* = F_{51}$,$(F_{44})^* = F_{52}$, $(F_{45})^* = F_{60}$, $(F_{46})^* = F_{56}$, $(F_{47})^* = F_{58}$, $(F_{48})^* = F_{57}$, $(F_{49})^* = F_{59}$, $(F_{50})^* = F_{55}$.

2 Results

There exist 61 non-isomorphic groupoids of order 3 with identity F_1 $(xy \cdot zx) = (xy \cdot z)x$ from possible 314 groupoids.

There exist 40 non-isomorphic groupoids of order 3 with identity F_2 $xy \cdot zx = (x \cdot yz)x$ (middle Moufang) from possible 196 groupoids.

There exist 61 non-isomorphic groupoids of order 3 with identity F_3 $(xy \cdot zx) = x(y \cdot zx)$ from possible 314 groupoids.

There exist 40 non-isomorphic groupoids of order 3 with identity F_4 $xy \cdot zx = (x \cdot yz)x$ (middle Moufang) from possible 196 groupoids.

There exist 158 non-isomorphic groupoids of order 3 with identity F_5 $(xy \cdot z)x = (x \cdot yz)x$ from possible 874 groupoids.

There exist 49 non-isomorphic groupoids of order 3 with identity F_6 $(xy \cdot z)x = x(y \cdot zx)$ (extra identity) from possible 239 groupoids.

There exist 61 non-isomorphic groupoids of order 3 with identity F_7, $(xy \cdot z)x = x(yz \cdot x))$ from possible 305 groupoids.

There exist 61 non-isomorphic groupoids of order 3 with identity F_8, $(x \cdot yz)x = x(y \cdot zx)$ from possible 305 groupoids.

2.1 About algorithms

We use standard approach to generation of groupods with identities. After that we generate all isomorphic images of every groupoid with

an identity of Bol-Moufang and then choose non-isomorphic groupoids. The last stage is the most time consuming.

3 Conclusion

In this paper we continue counting non-isomorphic groupoids of order three with some Bol-Moufang identities.

Acknowledgments. We were supported in the frame of projects 20. 80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data"and "Algebraic systems with additional structures, theory quasigroups and their application in coding theory".

References

[1] F. Fenyves. *Extra loops. II. On loops with identities of Bol-Moufang type*, Publ. Math. Debrecen, vol. 16, pp. 187–192, 1969.

[2] Vladimir Chernov, Alexander Moldovyan, and Victor Shcherbacov. *On some groupoids of order three with Bol-Moufang type of identities*. In: Proceedings of the Conference on Mathematical Foundations of Informatics MFOI 2018, July 2-6, 2018, Chisinau, (Chisinau, Moldova), 2018, pp. 17–20.

[3] V. D. Chernov, N.N. Malyutina, and V. A. Shcherbacov. *Groupoids of order three with Bol-Moufang identities up to isomorphisms*. In: Abstracts of International conference Mathematics & IT: Research and Education (MITRE-2021) Moldova State University July 01 - 03, 2021, Chisinau, Republic of Moldova, p. 24–25.

[4] Grigorii Horosh, Victor Shcherbacov, Alexandru Tcachenco, and Tatiana Yatsko. *On groupoids with Bol-Moufang type identities*, Computer Science Journal of Moldova, vol.28, no.3(84), 2020, pp. 314–327.

[5] K. Kunen. *Quasigroups, loops and associative laws*, J. Algebra, 185, no. 1, 1996, p. 194–204.

Vladimir Chernov[1],
[1] Researcher /Shevchenko Transnistria State University
E–mail: volodya.black@gmail.com

Valentina Demidova[2]
[2] Researcher/Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mail: valentina.demidova@math.md

Nadeghda Malyutina[3]
[3] Ph.D. Student/Moldova State University
E–mail: 231003.bab.nadezhda@mail.ru

Victor Shcherbacov[4]
[4] Researcher/Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mail: victor.scerbacov@math.md

Proceedings of Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

Considerations on the Artificial Intelligence Strategies

Svetlana Cojocaru, Constantin Gaindric, Tatiana Verlan

Abstract

The paper examines the policies of European countries regarding the use and development of applications based on Artificial Intelligence. Along with the examination of the strategic documents of the European Commission, some of them are studied at national level. Also, the situation in this field in the Republic of Moldova is analyzed.

Keywords: Artificial intelligence, strategy, education, research and development.

1 Introduction

Artificial intelligence (AI) becomes the driving force of the digital age. AI-based applications are becoming more frequently used, often without this being explicitly realized. Automatic translation, the quality of which is getting better, and also contextual advertising are just two of such examples, which each of us knows. Among the objectives of the new Digital Europe 2021-2027 Program, there is the massive implementation of solutions based on artificial intelligence, especially in critical areas such as climate change or health.

In Section 1 of this article, we will examine the evolution of the definition of artificial intelligence, as well as the basics of the strategy adopted by the European Union. Community countries, in turn, have developed their own national strategies (or they are in the process of developing them). In Section 2 we will examine the specifics of the

approaches in three countries: Estonia, Bulgaria, and Romania. The topic of Section 3 is the situation in the Republic of Moldova examined from the point of view of human potential, infrastructure, existing developments that could be the subject of public and private sector implementations, as well as the reflection analysis of the usage aspects of solutions based on artificial intelligence in the government program, adopted in August 2021.

2 Review of definitions of the term "Artificial Intelligence"

Before considering approaches to the problem of elaboration and employing systems of Artificial Intelligence (AI), let us make a small review of definitions of the term "Artificial Intelligence". We should note that it is considered to be used for the first time by John McCarthy in 1956 at a summer seminar at Dartmouth College (Hanover, USA). Later, in 2004, in his article [1] with questions of a layman about AI, he gives some interesting and explaining answers. Here are several of the basic questions and answers:

"**Q.** What is artificial intelligence?
A. It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.

Q. Yes, but what is intelligence?
A. Intelligence is the computational part of the ability to achieve goals in the world. Varying kinds and degrees of intelligence occur in people, many animals and some machines.

Q. Isn't there a solid definition of intelligence that doesn't depend on relating it to human intelligence?

A. Not yet. The problem is that we cannot yet characterize in general what kinds of computational procedures we want to call intelligent. We understand some of the mechanisms of intelligence and not others.

Q. Is intelligence a single thing so that one can ask a yes or no question "Is this machine intelligent or not?"?
A. No. Intelligence involves mechanisms, and AI research has discovered how to make computers carry out some of them and not others. If doing a task requires only mechanisms that are well understood today, computer programs can give very impressive performances on these tasks. Such programs should be considered "somewhat intelligent"."

There are a lot of definitions of the term "Artificial Intelligence" in different kinds of literature that to a varying degree explain its essence.

The compilers of the explanatory dictionary on artificial intelligence tried to collect and systematize special terminology on artificial intelligence and intelligent systems [2]. So, they define AI in two parts:

"1. A scientific direction, within the framework of which the problems of hardware or software modeling of those types of human activity that are traditionally considered to be intellectual are set and solved.

2. The property of intelligent systems to perform functions (creative), which are traditionally considered the prerogative of a person. Intelligent System (IS) - a technical or software system capable of solving problems traditionally considered creative, belonging to a specific subject area, knowledge about which is stored in the memory of IS. The structure of IS includes three main blocks - knowledge base, solver and intelligent interface."

Other several definitions of AI, but not all, are as follows:

- Artificial intelligence is the use of computers and systems for the simulation of the human mental process to solve problems and make decisions [3].

- Science called "artificial intelligence" is included in the complex of computer science, and the technologies created on its basis belong to information technologies. The task of this science is to provide reasonable discourses and actions using computing systems and other artificial devices [4].

- Artificial intelligence is a technology, or rather a direction of modern science, which studies ways to train a computer, robotic technology, and an analytical system to think intelligently like a person. The main goals of AI:

  - Creation of analytical systems that have intelligent behavior, can independently or under the supervision of a person learn, make predictions and build hypotheses based on the data set.
  - Implementation of human intelligence in a machine - the creation of robotic assistants that can behave like humans: think, learn, understand, and perform assigned tasks [5].

- Artificial intelligence (AI) is the foundation for simulating human intelligence processes by creating and applying algorithms embedded in dynamic computing environments. Simply put, AI is trying to make computers think and act the way humans do. Achieving this goal requires three key components: Computing systems, Data and data management, Advanced AI algorithms (code). The closer the desired result is to humans, the more data and processing power is required [6].

- Artificial intelligence is defined as a complex of technological solutions that allows simulating human cognitive functions, including self-learning and finding solutions without a predetermined algorithm, as well as obtaining results, when performing tasks, comparable, at least, to the results of human intellectual activity [7].

When analyzing, these definitions, as the reader can conclude, are not controversial. Rather, they complement each other to provide a

clearer picture. The comprehensive definition is given by the independent High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission [8]:

> "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans3 that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors, and actuators, as well as the integration of all other techniques into cyber-physical systems)."

The AI HLEG proposes to use this definition when projecting strategy for AI development.

The White Paper on AI created by the European Commission (EC) emphasizes the fast development of AI in modern society. This fact permits us to change the working style in different fields of our life. Financial and banking systems, industry, education, online trade, politics, healthcare, agriculture, household equipment – these are only some directions, where AI usage facilitates the respective processes.

Nevertheless, there are also serious risks to which AI usage can give rise, "such as opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes" [9]. So, the EC maintains regulatory measures and funding possibilities for a dual purpose: on the one hand, to promote the AI implementation; on the other hand, to pay attention to potential risks

that this new technology may bring: "This White Paper presents policy options to enable a trustworthy and secure development of AI in Europe, in full respect of the values and rights of EU citizens."

Also, there is another regulatory document of the European Commission, Coordinated Plan on Artificial Intelligence, which "puts forward a concrete set of joint actions for the European Commission and Member States on how to create EU global leadership on trustworthy AI" [10]. This document indicates, first of all, three main conditions for achieving the objectives set by Coordinated Plan: 1) appropriate governance and coordination framework; 2) data (large, high-quality, secure, and robust datasets); 3) computation infrastructure (necessary for storing, analyzing, and processing the increasingly large volumes of data).

3 Case study: approaches to the artificial intelligence use in different European countries

In this section, we will examine the examples of implementation of the Coordinated plan on artificial intelligence (2021) in several EU countries. Multiple international reports [11-13] analyze state of the art in the field of national AI strategies in terms of the following policy areas:

- **Human capital.** This section includes policies aimed at educating the population of all ages in the field of using and developing AI-based solutions. They cover both the respective courses in the programs of educational institutions of all levels, as well as refresher trainings for employees of different specialties to cultivate their skills needed to operate with AI-based systems.

- **From the lab to the market** policies are inclined to support research and innovation in AI in order to assure business growth in the private sector and increase the efficiency of public services.

- **Networking** refers to collaborations in the field of AI promoted by the private and public sectors, including at the international

level.

- **Regulation** covers policies related to ethical issues, the regulatory framework, the adoption of international standards.
- **Infrastructure** refers both to the development aspects of digital and telecommunications infrastructure itself and to solving the problems of data collection, use, and sharing at the national and international level.

In our examples, we will refer mainly to the first two aspects.

The first example is that of Estonia. We chose this country for our case study for two reasons:

- its success in digital transformation is well-known;
- it is part of the list of ex-Soviet countries, so we could say that after the declaration of independence in August 1991, the Republic of Estonia and the Republic of Moldova were at the same starting point.

The Government of Estonia adopted its National Artificial Intelligence Strategy in July 2019. The document [14] envisaged the implementation of a series of actions divided into four compartments:

1. Advancing the uptake of AI in the public sector in Estonia;
2. Advancing the uptake of AI in the private sector in Estonia;
3. Developing AI R&D and education in Estonia;
4. Developing a legal environment for the uptake of AI.

Most actions (30 in number) are provided concerning the public sector. Along with those of information, involvement of public authorities, trainings, etc. we will also mention the specific presence in this compartment of the R&D projects related to the implementation of automatic AI-based decision-making support in Estonian state institutions. Within this project [15] several applications based on artificial intelligence approaches have been developed, including:

- A system to support NEET youth, which is made available to municipal workers in the field of social assistance, especially those dealing with child protection and youth problems, to identify and support young people aged 16-26 who are Not in Education, Employment, or Training.

- Prediction model for the healthcare needs of patients with chronic illnesses. The system, implemented in the pilot version and based on machine learning algorithms, is made available to family physicians by assisting in identifying patients on their list with multiple chronic illnesses who would benefit most from additional help with prevention, counseling, and follow-up care to improve their quality of life.

- Machine learning software to match job seekers with employers. Basing on the European Skills, Competences, Qualifications, and Occupations (ESCO) classification system, developed by the European Commission, which defines skills needed in many areas of life, there was elaborated the machine learning algorithm which chooses candidates with skill categories suitable for the corresponding job profile. At the time of reporting (2020), the system operated with over 400,000 user profiles (remember that the population of Estonia is 1 million 325 thousand inhabitants) and a smaller number of workplace profiles, and the hiring process could be fully automated and did not take more than 5-10 days.

- Machine vision AI solution for better traffic management. The system is implemented in Tallinn and serves to monitor road traffic, especially public transport. Based on the information collected and processed within the system, it is possible, along with other aspects, to make decisions about parking problems or road construction.

We will also mention the special R&D actions in the field of R&D, within which three relevant research groups are funded: AI and machine learning, data science and big data, robot-human cooperation. For the

research and elaborations carried out by these groups, 1.5M EUR were allocated annually.

A series of activities are also envisaged in the field of education, the leader being the University of Tartu. An important role belongs to the IT Academy program (English brand name StudyITin.ee), which includes the collaboration of the state, educational institutions, and ICT companies to ensure an advanced quality of studies and research in the field of ICT. It involves training about 50 master's students specializing in AI at the University of Tartu in the period 2020-2023, reviewing the curricula of courses in general schools with the inclusion of AI subjects.

Last but not least, we would like to point out that the AI Program in Estonia is called the "Kratt plan". Kratt is a character from local mythology, an artificial creature, who serves his master by performing various works, which the master orders him to do. The need to pay tribute to the devil for the creature to function also alludes to some ethical issues, which are intensely discussed concerning the vast application of AI in various fields.

From the above, we can conclude that this country has all the chances to achieve its ambitious goal formulated as follows: "Estonia could become the role model and testbed for the rest of the world as a place where Kratt, or AI, is put to work for the people's well-being in both the public and private sectors" [16].

Another example, which we will examine, is that of Bulgaria. We have selected this country thanks to several similarities, which we can observe, as we will see from the following, as compared with the situation in the Republic of Moldova. The country's strategic document is entitled "Artificial intelligence for intelligent growth and a prosperous democratic society", the project was developed by the Bulgarian Academy of Sciences and approved in December 2020. The concept envisages the realization of a series of actions over the next ten years, based on existing results in the field of artificial intelligence uptake and the development of AI-based applications [17].

Several sources (such as [18]) mention the existence of over 50 com-

panies working with artificial intelligence applications, the most important areas in this regard being retail, finance, and media. The research and elaborations in the field of natural language processing also have international recognition. In [19] it is emphasized that the main areas of specialization in Bulgaria are big data, predictive analytics, data science, and chatbots. According to the same source, the share of the IT sector in GDP formation in Bulgaria is 3.4%, which is very close to the figures in Moldova: according to data from 2015-2019, the IT industry has reached 3.1% of GDP [20].

The strategy document stipulates that a fundamental proposal for Bulgaria is "focusing on technological specialization in the field of data economy, as the country would have difficulty when realizing strong industrial specialization due to the lack of a critical mass of top industrial companies in the AI sector. Today, the trend is for data to come to the fore in AI and for the emphasis on automatic self-learning to shift from algorithms to data" [17].

The following domains and directions for AI development and implementation are established as the priority ones:

- Software industry;
- Creating AI applications for educational purposes;
- AI applications in public services;
- Intelligent agriculture;
- Applications of AI in healthcare and medicine;
- Applications of AI in ecology and environment.

Special attention in the strategic concept is paid to the education and research domains, namely here we find several features, specific to our country too. Whereas work in the IT sector is much better paid than research, most young people either do not want to start or begin and leave their careers in universities and research institutes, preferring to work in IT companies in the country or abroad.

The basic recommendation for Bulgaria is the need to overcome the fragmentation between small units that develop AI and creation

of the conditions for building human potential in a connected national academic environment. Thus, it is proposed:

- Establishment of a Bulgarian center of excellence in AI, which will unite scientific organizations and universities with proven achievements in the field of AI research;
- Involvement of Bulgarian research teams in European artificial intelligence and digitization networks;
- Inclusion of Bulgarian research teams in European testing and experimentation centers related to healthcare, robotics, and agriculture.

In the case of Romania, it is also necessary to mention the initiative of the researchers from the Romanian Academy, who in October 2019 published a document entitled "Manifesto for adaptation to the digital age" [21]. The document specifies the favorable factors for Romania, namely:

- the weight of the ICT sector in GDP (according to the 2019 Country Report, published by the European Commission [22], it was 6-7%);
- Internet infrastructure with high traffic speeds;
- wide penetration of mobile devices among the population;
- the special receptivity of young people towards these technologies.

In the recommendations elaborated by the authors of the "Manifesto" several directions of action are established, in which an important role belongs to research, education, public administration, media.

Although Romania does not yet have a finalized national strategy (in [12] it is mentioned that its elaboration is "in progress"), the Romanian Digitization Authority (RDA) considers that "Artificial Intelligence can revolutionize the activity of public administration", thereby offering "better public services, safer transport systems, personalized products and services that are cheaper and more sustainable" [23]. On July

14, 2021, RDA initiated public consultations in the project "Strategic framework for the adoption and use of innovative technologies in public administration 2021-2027 – solutions for activities efficiency", the purpose of which is to formulate the position of the academic environment, of research-development-innovation sector, of the central and local public administration authorities and of the private environment concerning the field of Artificial Intelligence. According to the communiqué, distributed by RDA, "the project will determine the national strategic framework and financial instruments for Romania's participation in European networks and ecosystems of scientific data management – open science, for the adoption of blockchain technologies in key areas of government intervention, such as identifying people and document management in the field of education, but also for outlining the national framework in the field of artificial intelligence, a technology that – through its complex applications – has a major impact on improving the quality of citizens' life and business development" [23].

4 Some aspects of promoting AI in the Republic of Moldova

The importance of information technologies has been realized in the Republic of Moldova since the 90s of the last century. Our country was among those that in 1990 included in the structure of its Government a Ministry of Informatics, Information, and Telecommunications. Subsequently, it underwent several changes in both name and duties, but regardless of them, ICT development and implementation policies were constantly promoted, thanks to which it was possible to create a communication infrastructure based on optic fibers, which had good coverage in the country; there have been implemented services intended for citizens and economic agents, based on digital technologies; measures have been carried out to equip schools with computers and connect them to the Internet (Program SALT, adopted in 2004, assumed the maintenance of physical access to the Internet for all schools of the country [24]).

Today, according to the analysis of the ICT sector involvement in the economy of the Eastern Partnership countries, carried out by the German Economic Team [25], in the respective sectors of Armenia, Belarus, Georgia, and Ukraine, the ICT revenues in 2019 accounted for 7.1% of GDP. In Moldova, there is observed a share of 5% of GDP, and exports of ICT services represent more than 15% of services exports and about 6.5% of total exports, but only 2% of Moldova's GDP, while ICT infrastructure is very well developed and the number of users with broadband internet access has increased significantly from 42% in 2014 to 117% in 2020.

The conditions for the IT sector development are good, because there is a developed infrastructure and the population uses ICT technologies extensively, and the companies' expenses are increasing according to the National Bureau of Statistics from 500 thousand lei in 2013 to 2500 thousand lei in 2019.

If in 2005 the index of internet penetration in Moldova was 7.4% compared to 35.5% in Europe, currently according to "The future of IT Landscape Report. The ultimate guide for IT buyers and investors looking to source in emerging Europe", developed by Emerging Europe in 2021 [26], the internet penetration rate in Moldova is already 76% with an increase of more than 10 times compared to 2005, while in Europe it is 88.2% with an increase of 2.48 times.

Thanks to higher salaries, the opportunities offered by ICT are attractive, and yet the share of the workforce in the ICT sector is relatively small.

Based on the data presented, let's see how the field of AI in the Republic of Moldova is presented in the light of EU regulatory documents. We will mention that by June 2021, 20 EU Member States and Norway had published their national AI strategies, while 7 Member States were in the final drafting phase. The Republic of Moldova adopted in October 2013 the "Digital Moldova 2020" strategy, which envisaged the actions within three pillars, namely: 1) Pillar I: Infrastructure and access – improving connectivity and network access; 2) Pillar II: Digital content and electronic services – promoting the generation of digital

content and services; 3) Pillar III: Capacities and use – consolidation of digital literacy and skills to allow innovation and stimulate usage. By the date of this article's preparation, no document has been made public, which would sum up the successes and failures in this strategy implementation.

On August 4, 2021, the Parliament of the Republic of Moldova approved the Government Activity Program "Moldova of Good Times" [27], which also contains a section dedicated to digital transformation. In this section as well as in others, the notions of artificial intelligence, intelligent municipality, intelligent instruments, etc. are encountered. The Program stipulates that "The state must be able to capitalize on the opportunities offered by the digital revolution, but also to manage the risks generated by it". However, the program does not explicitly state, as in other countries' policy documents, that artificial intelligence will influence the increase of efficiency of services provided to citizens by state authorities, relaunch industry, streamline agriculture, mitigate climate change, and improve healthcare. The provisions of the program are limited to "Studying and exploitation of initiatives and programs of EU countries in the field of adopting artificial intelligence technologies, robotics, blockchain, smart contracts and other emerging technologies to modernize public and private digital infrastructures with the purpose to deliver better services, operational effectiveness, and strengthening of the country's cybernetic capacity".

Thus, the government intends to use modern working tools by intensifying the application of information technologies to exclude the flow of paper documents in administrative processes. On the other hand, the White Paper [9] rightly states: "Europe's current and future sustainable economic growth and societal wellbeing increasingly draws on value created by data. AI is one of the most important applications of the data economy". In this context, the digital inclusion of local authorities is envisaged by creating a digital platform with access to centralized information resources, and public services to be rethought and modernized with a focus on the citizen.

The continuation of the Government Services' modernization, tak-

ing into account the Government's vision expressed in the Program, based on the Action Plan concerning public service modernization reform, would capitalize on and continue the achievements reached in the framework of the ongoing e-Government Transformation Project. Also, it would contribute to a) reorganizing public administrative services for the purpose to be provided implicitly and electronically as a priority, with the result of the service delivered in the form of an electronic document; b) increasing access, efficiency, and quality in the provision of government services. A key element of success is the evaluation by beneficiaries – citizens - of the quality and accessibility of services because a considerable part of citizens does not trust the quality and safety of electronic services. And another aspect is the use of the set of artificial intelligence technologies that combine data, algorithms, and ascending dynamics of internet penetration.

To have an overall view of national policy initiatives and national AI strategies, we will return to their examination in the light of the five policy domains, presented in the previous section:

- **Human capital.** There are currently five universities in the Republic of Moldova, where ICT specialists are being educated, but there is no master's program specifically aimed at AI.
- Ways and means of **passing from the laboratory to the market**. As it was mentioned earlier, this compartment includes policy initiatives to encourage research and innovation in AI for business growth in the private sector and public services' efficiency increase. Tools are also included to facilitate the experimentation of AI pilot products and newly developed services.

The propulsion of new AI products from the laboratory to the market can only succeed in an enterprise-based environment, with funding for research and innovation in AI, which would support the transformation of AI concepts into successful products and services. Mechanisms are needed for the adoption and use of AI in public administration. In this context, EU countries have taken measures to stimulate AI research and have developed or are in the process of setting up national centers

of competence in AI research. Some centers are aimed at many domains of research in AI, others are focused on autonomous systems, cyber security procedures for AI systems, machine learning, Data Science. A key priority in AI research is a collaboration with research centers in different countries. The most frequently reported sectors in national strategies are production, agriculture, healthcare, transport, and power engineering.

We will make a brief overview of the Republic of Moldova's research related to AI. From the analysis of different countries' national strategies, we can see that several of them have given priority to AI language technologies for interactive dialogue systems and virtual assistants for personalizing public services. Denmark, Norway, Portugal, Slovakia, and Spain have included support policies for research in natural language processing. The research carried out in the Republic of Moldova in natural language processing is in line with EU visions, and the results obtained over the years are at the European level. In particular, we will emphasize the achievements in the recovery of the country's cultural heritage. The systems for digitizing old texts (starting with the 17th century), developed within the "Vladimir Andrunachievici Institute of Mathematics and Computer Science", allow the restoration of works of historical value in the wide circuit, offering specialists in various fields and the general public as well a tool for accessing these works in a convenient, editable format, in an original or contemporary script.

The project "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" is in the trend of EU-recommended research. Several national projects and those of the partnership with EU countries have aimed to develop medical information systems oriented to providing physicians with support in the diagnostic process, as well as modeling the course of patient treatment. The formation of a Network for informational methods in supporting persons predisposed to preventable strokes using common devices has been initiated. The international project "An Adaptive Decision Support Framework for the Management of Mass Casualty via an Artificial Intelligence

Based Multilayered Approach integrating an Intelligent Reachback Information System" is underway, in which, together with the Republic of Moldova team, researchers from Germany, Croatia, Romania, USA participate. Researchers from our country have joined some COST Action projects, such as "European Network for Combining Language Learning with Crowdsourcing Technique", "A Network for Gravitational Waves, Geophysics and Machine Learning", etc.

The problem of AI is of importance that should not be neglected, a fact that requires the creation of a national center of competence in AI research, because the number of researchers in research institutes and universities is far below the limit of necessities, being worthy of following the experience of several EU (but not only) countries, which directed AI research in national programs to the needs of countries, without neglecting the general perspective directions.

The innovation and usage of AI in public administration are stimulated, including AI programs for public services, e-government strategies to improve the digitization of public administration processes, public procurement, and exchange of good practices. The e-government center at the Government of the Republic of Moldova managed to offer the society a series of services, the last being MDelivery, which significantly changed the processes and time to obtain some documents for which citizens previously lost time and effort, but also, which is more important, have to some extent influenced the attitude of the population towards the government.

For the time being, a greater interest in the application of AI-based technologies in our country is attested in the finance & banking and insurance sectors. The National Bank of Moldova is among the institutions that, operating with such systems, will have the possibility of interconnected supervision of all banking operations, as it will function as a single point of access and visualization of information contained in numerous databases. The concept of operation is based on risk analysis, which will make it possible to detect suspicious activities and issue alerts, early identification of risks of money laundering and terrorist financing, and the detection of suspicious changes in the ownership

structure of banks [28].

- **Networking** includes the set of virtual means and AI collaboration initiatives in the private and public sector, including with foreign people and companies. Networking includes dissemination policies, promotional campaigns, and mapping of AI applications. Many governments have put in place policies to build innovation communities by bringing together technology companies, research centers, and innovation actors. Many countries also set policies to attract skills and investment from AI abroad. In this respect, some countries have dedicated strategies, such as the researcher mobility program in Cyprus and the future Spanish Talent Hub program. Other policies aim to improve working conditions for foreign talent by facilitating administrative procedures. The Czech Republic, Finland, Italy, Malta, Portugal, and Spain are implementing this through starting visas and fast services for valuable talent coming from abroad.

  Our country has other problems, namely the loss of young specialists who either emigrate to countries that provide them with a well-paid and interesting job or work in foreign companies, which have other objectives than their home country.

Most countries exploit social channels to raise awareness in respect of AI and increase networking opportunities. Slovenia intends to launch a communication platform for the collection and dissemination of good practices and case studies on the use and implementation of AI in society. Hungary announces the annual award for innovations and AI application projects. Therefore, the Republic of Moldova can also take over a series of good practices in this area.

- **Regulations** provide policies that address issues that refer to human rights, confidentiality, fairness, algorithmic prejudice, transparency and explicability, security and responsibility, etc. To facilitate the development of ethical guidelines, many governments have formed AI ethics committees or councils. These bodies are

tasked with developing recommendations on ethical problems and monitoring the use and development of AI technologies. Slovakia is preparing a new act in respect of data to better define data protection regulations, data access principles, and open data regulations. Finland and Portugal are developing national regulations for determining liability problems.

- **The infrastructure** focuses on the problems of digital and telecommunications infrastructure development and provides initiatives to encourage data collection, use, and sharing. Since AI algorithms imply large amounts of data, it is crucial to establish an environment conducive to infrastructure development to ensure reliable, high-quality data that can be shared with users in an accessible and robust way.

Several EU Member States have drawn up national strategies to lay the foundations for the use and exchange of data that describe the actions needed for open data governance, the creation of data warehouses, the improvement of data interoperability, and the protection of individual and collective rights.

Open data platforms and portals have been developed in all EU Member States, Norway, and Switzerland. They usually aim to provide free access to public administration data.

In this direction, vigorous measures are required from the e-government service of the Government of the Republic of Moldova.

These policy areas are in line with the actions proposed in the Coordinated Plan on Artificial Intelligence [10] and with the policy recommendations addressed to governments, contained in the OECD Recommendation on AI [29].

5 Conclusion

The European Union, as the main goal, has proposed massive implementation of digital technologies in enterprises, putting them at the service of citizens and public administrations. Analyzing the policy

documents related to AI in different countries, we note that an important role is assigned to education and research & development. Less visible this aspect appears in the program documents of the Republic of Moldova. Along with the actions envisaged for study and exploitation of EU countries' initiatives and programs on artificial intelligence, as well as the interaction with the Ad hoc Committee on Artificial Intelligence of the Council of Europe, the study on national strategies in other countries indicates the need for active involvement of researchers, but also their support from the state for the achievement of a comprehensive and efficient digital transformation.

Acknowledgments. The research was supported by the project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge, and big data".

References

[1] J. McCarthy. *What is Artificial Intelligence?*, Computer Science Department Stanford University, Stanford, 2004 Nov 24. https://homes.di.unimi.it/borghese/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_whatisai.pdf.

[2] A. N. Averkin, M. G. Gaaze-Rapoport, D. A. Pospelov. *Explanatory Dictionary of Artificial Intelligence*, Moscow: Radio and communication, 1992, 256 p. http://www.raai.org/library/tolk/aivoc.html#L208.

[3] IBM Cloud Education. *Artificial Intelligence*, June 3, 2020, https://www.ibm.com/ru-ru/cloud/learn/what-is-artificial-intelligence.

[4] G. S. Osipov. *Artificial Intelligence: State of Research and Future Outlook*. Artificial Intelligence News, no. 1, 2001. http://www.raai.org/about/persons/osipov/pages/ai/ai.html.

[5] Calltouch. *Artificial Intelligence*, 2021, https://www.calltouch.ru/glossary/iskusstvennyy-intellekt/.

[6] NetApp. *What is Artificial Intelligence?*, https://www.netapp.com/ru/artificial-intelligence/what-is-artificial-intelligence/.

[7] *National Strategy for the Development of Artificial Intelligence.* TAdviser, 2021/08/02 19:12:40, https://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%9D%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D1%81%D1%82%D1%80%D0%B0%D1%82%D0%B5%D0%B3%D0%B8%D1%8F_%D1%80%D0%B0%D0%B7%D0%B2%D0%B8%D1%82%D0%B8%D1%8F_%D0%B8%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE_%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D0%B0.

[8] Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. *A Definition of AI: Main Capabilities and Disciplines*, 8 April 2019.

[9] European Commission. *White Paper On Artificial Intelligence – A European approach to excellence and trust*, Brussels, 19.2.2020.

[10] European Commission. *Coordinated Plan on Artificial Intelligence 2021 Review*, Brussels, 21.4.2021.

[11] V. Van Roy. *AI Watch – National strategies on Artificial Intelligence: A European perspective in 2019*, EUR 30102 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-16409-8, doi: 10.2760/602843, JRC119974.

[12] V. Van Roy, F. Rossetti, K. Perset, and L. Galindo-Romero. *AI Watch – National strategies on Artificial Intelligence: A European perspective*, 2021 edition, EUR 30745 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-39081-7, doi: 10.2760/069178, JRC122684.

[13] *State of implementation of the OECD AI Principles: Insights from national AI policies*, OECD Digital Economy Papers, No. 311, OECD Publishing, 2021, Paris, https://doi.org/10.1787/1cd40c44-en.

[14] *Estonia's national artificial intelligence strategy 2019-2021*, https://f98cc689-5814-47ec-86b3-db505a7c3978.filesusr.com/ugd/7df26f_27a618cb80a648c38be427194affa2f3.pdf.

[15] Fabio Chiusi, Sarah Fischer, Nicolas Kayser-Bril, and Matthias Spielkamp (eds.). *Automating Society Report 2020*, AlgorithmWatch gGmbH, Berlin, Germany, 298 p. https://automatingsociety.algorithmwatch.org/.

[16] *National AI strategy for 2019-2021 gets a kick-off.* https://e-estonia.com/nationa-ai-strategy/.

[17] *Development concept of artificial intelligence in Bulgaria until 2030: Artificial intelligence for smart growth and a prosperous democratic society.* https://www.mtitc.government.bg/sites/default/files/koncepciyazarazvitienaiivbulgariyado2030.pdf. (in Bulgarian).

[18] *Artificial intelligence ecosystem in Bulgaria.* https://investsofia.com/wp-content/uploads/2019/06/Artificial_intelligence_ecosyste m_in_Bulgaria_2019-SeeNews_and_Vangavis.pdf.

[19] G. Angelova et al. *Role of education and research for artificial intelligence development in Bulgaria until 2030.* In: Mathematics and Education in Mathematics. Proceedings of the Fiftieth Spring Conference of the Union of Bulgarian Mathematicians, 2021, pp.71–82.

[20] The ICT sector has become one of the locomotives of economic growth in the Republic of Moldova. https://www.moldpres.md/news/2020/09/03/20007029. (in Romanian).

[21] *Manifesto for adaptation to the digital age.* https://acad.ro/mediaAR/com2019/c1016-ManifestEraDigitala.htm. (in Romanian).

[22] *Country Report Romania 2019 Including an In-Depth Review on the prevention and correction of macroeconomic imbalances.* https://ec.europa.eu/info/sites/default/files/file_import/2019-european-semester-country-report-romania_en.pdf.

[23] *ADR launches the public consultation process on the use of artificial intelligence in Romania.* https://www.adr.gov.ro/adr-

lanseaza-procesul-de-consultare-publica-privind-utilizarea-inteligentei-artificiale-in-romania/. (in Romanian).

[24] L. Burtseva et al. *Digital divide: A glance at the problem in Moldova.* In: Information Communication Technologies: Concepts, Methodologies, Tools, and Applications. IGI Global, Vol. IV, 2008, pp. 2531–2565.

[25] Carolin Busch, Ricardo Giucci. *Moldova's ICT sector in comparison to Ukraine, Belarus, Georgia and Armenia, 2021.* https://www.german-economic-team.com/moldau/wp-content/uploads/sites/4/GET_MDA_PB_04_2021_EN.pdf.

[26] https://emerging-europe.com/it-landscape-report/.

[27] https://unpaspentru.md/2021/08/03/program-de-activitate-al-guvernului-moldova-vremurilor-bune/.

[28] https://www.bnm.md/ro/content/premiera-regiune-moldova-va-implementa-cu-suportul-usaid-o-solutie-it-de-ultima-generatie.

[29] OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449, 2019. https://www.fsmb.org/siteassets/artificial-intelligence/pdfs/oecd-recommendation-on-ai-en.pdf.

Svetlana Cojocaru[1], Constantin Gaindric[2], Tatiana Verlan[3]

Vladimir Andrunachievici Institute of Mathematics and Computer Science,
5, Academiei street, Chisinau, Republic of Moldova, MD 2028

[1] E–mail: svetlana.cojocaru@math.md

[2] E–mail: constantin.gaindric@math.md

[3] E–mail: tatiana.verlan@math.md

Proceedings of Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

On XML Standards to Present Heterogeneous Data and Documents

Alexandru Colesnicov, Ludmila Malahov,
Svetlana Cojocaru, Lyudmila Burtseva

Abstract

The article discusses XML presentation of heterogeneous data on cultural heritage in the form of Romanian documents in the Cyrillic script. An example of such presentation is standards for health care records. Currently a suitable framework for development of necessary presentation is Europeana Data Model. The necessary adaptation of presentation standards are considered.

Keywords: cultural heritage, digitization, heterogeneous data, XML-based presentation standards.

1 Introduction

Old documents subject to digitization often contain heterogeneous content: text, images, musical scores, mathematical formulas, etc. The recognition of these documents produces not only text outputs and images in electronic form, but script presentations of other fragments [1]. The recognition of heterogeneous content needs specific recognition agents for each type of page fragment and generates several resulting files.

Storing and searching the recognized heterogeneous documents is a special challenge also.

An example of successful solution of these problems exists in the domain of medical documentation. XML is used as top-level structure for document presentation due to its inherent extensibility.

For cultural heritage objects, the XML-based Europeana Data Model (EMD) is a suitable base to define their corresponding description [2].

The paper at first presents our approach to the digitization of heterogeneous documents. Section 3 describes examples of standardized XML-presentation of heterogeneous data: US standards for health care records, and EMD that is the modern approach oriented to cultural heritage. Section 4 sums up ways to configure our data structures in the EMD framework.

2 Digitization of heterogeneous documents

Heterogeneous documents, or documents with heterogeneous content, contain fragments that obey certain formal rules. Examples are mathematical and chemical formulas, chess notation and diagrams, electronic circuits, etc. The main properties of such content: 1) it is not a text in a natural language nor pure image; 2) there is a scripting language for its description; 3) a graphical presentation can be reproduced from the script. During digitization we should obtain the script presentation of all such fragments in the document. We described the process of heterogeneous document digitization in [3].

The digitization is performed in 11 stages. It uses and produces the following types of source, intermediate, and *final (7 types of 10)* data.

A Graphical document
B *Page image in electronic form*
C *Page map*
D *Page fragment*
E *Script equivalent of a page fragment*
F *Extracted metadata*
G *Script equivalent of a page*
H *Reconstructed page*
I Verification log
J Error report

All final data may be used in further work with the digitized docu-

ment; they and all their parts should be stored and available for search, and even for execution. This makes the structure of the digitized document unexpectedly sophisticated.

The usual and obvious solution for the presentation of the digitized heterogeneous document is to base this presentation on XML.

3 Examples of standard presentation of heterogeneous data

An important practical experience of data standardization is accumulated in health care, especially in the USA [4].

Standards used across health care organizations fall into four large groups: terminology standards, content standards, data exchange or transport standards, and privacy and security standards.

Less attention is paid to standardization of data store formats. Standards are applied starting from the API level. If a standard API call returns data in a standard format, the internal representation of these data is unimportant.

Non-textual and non-image information is kept and transmitted encoded. There are a lot of standard codes: codes for diagnoses, codes for medical procedures, codes for all kinds of health-related services, codes for dental treatment, codes for clinical information, codes for lab orders and results, codes for pharmacy products, and codes for clinical drugs.

As to images, Digital Imaging and Communications in Medicine (DICOM) is an international standard supported by all medical devices. The DICOM file has the extension .DCM and contains metadata plus from 0 to 7 images.

Health care data transmission uses XML and JSON formats.

Our second example is EDM [2], a model developed in the frame of Europeana project. The web-portal of unified access to growing digital cultural resources, Europeana, was launched in 2009 as a result of massive digitization of European cultural heritage in 2000s. Its basis is the metadata standard Europeana Data Model that was announced

in 2010. The main aim of the EDM is to provide a simple workflow for adding any local collection to Europeana portal. Thus, EDM elements set can be extended by each new provider, as he joins the Europeana information space. Besides the new elements introduced by Europeana, EDM has the set of elements re-used from other namespaces. It is supported by clear documentation and schema checking tools. Thanks to such support, EDM has become the XML-based standard for cultural objects metadata representation.

EDM replaces and widely extends the first model, which was called European Semantic Elements (ESE). More exactly, ESE was "the lowest common denominator" of semantics used in different cultural heritage sectors, like museums, libraries, archives, and audiovisual collections. EDM reverses this approach and covers all community standards, for example, LIDO for museums, EAD for archives, and METS for libraries [5, p. 4–5].

Today, Europeana joins over 58 million cultural heritage items from around 4,000 institutions. The ascent to BigData level creates problems common for such huge collections. To facilitate Europeana management the portal was divided into two ones:

- **Europeana,** which deals with ready collections only, provides their store, search and metadata retrieving;
- **EuropenaPro,** which deals with technical infrastructure, develops and maintains technical solutions for showcasing, sharing and using digital cultural heritage. All products developed in the frame of EuropenaPro are free and mostly open sources. EuropenaPro has own folder github.com/europeana, where open sources solutions can be obtained.

Having a long history and significant results, Europeana does not intend to stop its development. In 2020 a new Europeana strategy (2020-2025) was issued [6]. It was declared, that in its further development Europeana will focus on supporting the digital transformation of European cultural heritage sector. The tasks and priorities for both collection management and technical solutions development were revealed.

4 Data structure configuration in the EMD framework

We see that the internal representation of the document is not standardized. Externally, only the metadata set and correspondence to data exchange standards are important.

Our project supposes the development of heterogeneous document representation based on XML. The metadata should correspond to the EDM that guarantees extensibility and flexibility.

Each digitized document should be kept as an archive containing all its outputs (texts, scripts, and images). The Open Office DOCX format gives us an example: it is a ZIP archive containing XML files with texts included inside XML and images. One of XML files describes the document structure and lists other files. XML files also contain metadata related to the document, for example, the author name.

We could adopt this structure. Scripts for specific non-textual content could be kept inside XML like usual texts but marked as to be rendered by specific agents that makes them similar to images.

Each type of content and each type of output have its specific set of metadata. For example, page image should be accompanied by its page number.

5 Conclusion

The study illustrates the possibility of adapting the existing models in order to solve the problem of heterogeneous data and documents presentation. Taking into account the fact that we operate with heterogeneous elements within the printed texts, the metadata spectrum can be reduced and connected to the types of components, specific to these texts. Keeping the processed document as an archive, containing all the final elements listed in Section 2, seems to be a plausible solution.

Acknowledgments. This work was prepared as part of the research project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”.

References

[1] A. Colesnicov, L. Malahov, S. Cojocaru, and L. Burtseva. *Semi-automated workow for recognition of printed documents with heterogeneous content.* Computer Science Journal of Moldova, vol. 28, no. 3(84), 2020, pp. 223–240.

[2] Available at: `http://pro.europeana.eu/edm-documentation`

[3] A. Colesnicov, S. Cojocaru, and L. Malahov. *Recognition of heterogeneous documents: problems and challenges.* In: Proceedings of the 5th Conference on Mathematical Foundations of Informatics, 3–6 July 2019, Iași, România, pp. 231–245. – Iași: Editura Universității "Alexandru Ion Cuza", 2019. – ISBN: 978–606–714–481–9.

[4] *Data Standards in Healthcare: Codes, Documents, and Exchange Formats.* October 23, 2020. Available at: `https://www.altexsoft.com/blog/data-standards-healthcare/`

[5] *EDM Primer.* Available at: `https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf`

[6] *Strategy (2020–2025). Empowering digital change.* DOI: 10.2759/524581. Available at: `https://pro.europeana.eu/files/Europeana_Professional/Publications/EU2020StrategyDigital_May2020.pdf`

Alexandru Colesnicov[1,2], Ludmila Malahov[1,3], Svetlana Cojocaru[1,4], Lyudmila Burtseva[1,5]

[1] "V. Andrunachievici" Institute of Mathematics and Computer Sc e, Chisinau, Republic of Moldova
[2] E–mail: `acolesnicov@gmx.com`
[3] E–mail: `ludmila.malahov@math.md`
[4] E–mail: `svetlana.cojocaru@math.md`
[5] E–mail: `luburtseva@gmail.com`

Punctilog: a New Method of Sentence Structure Representation

Ioachim Drugus, Tudor Bumbu, Victoria Bobicev, Victor Didic, Alina Burduja, Alexandr Petrachi, Victoria Alexei

Abstract

We present the experiments on sentence syntactic structure re-codification from dependency grammar to punctilog, a novel methodology of sentence structure representation. Our goal is to create a corpus annotated using this convention; to this end, we re-use the corpora already created by the Universal Dependency project. Several algorithms had been developed for the transformation. We discuss the obtained structures and frame the necessary steps to improve the results.

Keywords: computational linguistics, sentence structure, dependency grammar, punctilog.

1. Introduction

Over the years, numerous methodologies have been used to represent sentence structures in computational linguistics. All these methodologies aimed to capture the most important component: the meaning of the sentence. Some of them were used largely in various projects; some are less known in the research community.

In this paper, we present the work on punctilog, yet another theory of sentence structure representation. Our goal is to create a corpus annotated using this convention; we plan to use the corpora already created by the Universal Dependency project. We developed several algorithms for the transformation and tested them on a small subset of the Romanian corpus. The results of the algorithms are quite different; we discuss the obtained structures and the possible ways to improve the results.

The rest of the paper is organized in the following sections: Section 2 introduces our motivation; the novel methodology named punctilog is described in Section 3; universal dependencies project is presented in Section 4; the proposed re-annotation algorithms are introduced in Section 5; Section 6 presents and analyses the obtained results; Section 7 concludes with a discussion and future work.

2. Motivation

Since the appearance of Computational Linguistics, multiple methodologies have been proposed for the representation of the sentence structure starting with the famous Chomsky grammars [1], tree adjoining grammar [11], link grammar [12], head-driven grammar [10], dependency grammar [3], and others. All developed methodologies had one common problem: they failed to capture the meaning of the text. This work aims to develop a formalism that will help to capture and represent the meaning of a sentence in a logical way.

3. Punctilog

An annotation symbolism referenced as "Punctuation Markup Language" with the abbreviation "PML", was introduced in [6]. However, there is yet another symbolism "Prague markup language" [8] competing for the abbreviation "PML", and there are also other uses of this abbreviation in domains that are far from linguistics. Therefore, a new term "Punctilog" is introduced for the symbolism described in [6]; this term is alternatively used in the paper. The term "punctuation markup language" is still used for the class of all such markup languages but the "PML" acronym is being avoided. Punctilog is one of the markup languages of this class, the one specified in [6], and there can be currently or emerge later many other languages of this class.

The Punctilog markup language consistently uses several punctuation marks according to the defined strict semantics and provides an extension mechanism that allows adding to the language new annotation elements by specifying their semantics according to the format prescribed by the extension mechanism. Since the time when mathematical discourse started making part of the discourses in natural languages, the punctuation marks used by natural languages started being treated as a system that is

extending over time and that needs to be managed. The expected uses of Punctilog are:

(a) To serve as a symbolism for marking up various meanings of expressions in a text;

(b) To assist in the management of the punctuation systems of various languages.

A short account of Punctilog can be given by the example below taken from [6] of a sentence and its Punctilog annotation, which uses all Punctilog's punctuation marks:

„Let's go swimming!" called Ion Chistruiatu, but his team, Gicu and Mihai, was up to it already.

([Let's go swimming!] :((called: (Ion ::Chistruiatu))), (((his: <team>): <Gicu, Mihai>) :(((was :<up to>): it) :already))

There are three bracket types used for Punctilog annotation. Square brackets '[', ']' also called "hard brackets" are used to indicate "direct speech". Angle brackets '<', '>' also called "chevrons" are used to indicate an "individual". Round brackets, also called "parentheses", and in this paper also called "soft brackets" are used to indicate a "constituent".

The round brackets, the parentheses, are used in Punctilog annotation in order to visualize the constituent structure of the sentence [7]. A text may be "ambiguous", i.e. it may be treated as having several interpretations and, accordingly, several constituent structures. To disambiguate a text means to arrange the parentheses according to a certain pattern, a "parenthetical pattern". The term "soft brackets" sounds like an appropriate synonym for the Punctilog's parentheses since these can be arranged and rearranged in different manners to obtain many parenthetical patterns and to choose one or several which are considered the most correct.

A pair of parentheses correctly inserted in a text partially disambiguates the text. Obviously, for full disambiguation, the parentheses should be arranged in such a manner that each pair of balanced parentheses comprise a pair of constituents. The expression "(x

y)" where x and y are constituents can be treated as a binary operation. In Punctilog, this operation is called association operation.

4. Universal Dependencies

Universal Dependencies[1] is one of the largest current projects dedicated to syntactic analysis and creation of the treebanks for multiple languages [2]. Its initial aim was the development of the consistent treebank annotation for many languages, intending to facilitate multilingual parser development and cross-lingual learning. The annotation scheme is based on an evolution of Stanford universal dependencies [3] and Google universal part-of-speech tags [4]. During the last years, multiple researchers joined the project and 202 treebanks in this convention for 114 languages have been released by May 2021 [5].

The annotation is coded in conllu format[2]. This format organizes the whole text by a word per line; each word is accompanied by its id, lemma, part of speech code, the id of a headword, and dependency link label. Figure 1 presents the graphical visualization of the dependency annotation created by conllu format viewer[3].

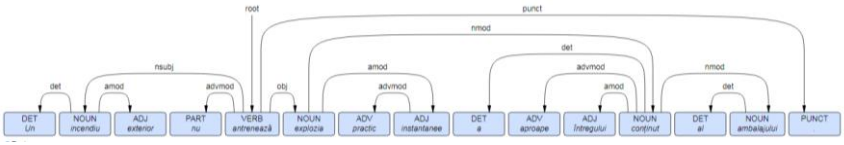

Figure 1. Graphical representation of the dependency annotation for the Romanian sentence "Un incendiu exterior nu antrenează explozia practic instantanee a aproape întregului conținut al ambalajului" (An external fire does not cause a virtually instantaneous explosion of almost the entire contents of the package).

5. Description of the Algorithms

There are several important differences between dependency grammar and punctilog:

[1] https://universaldependencies.org

[2] https://universaldependencies.org/format.html

[3] https://urd2.let.rug.nl/~kleiweg/conllu/

- dependency graph connects words as its nodes with dependency arcs; punctilog forms a tree graph with words as their leaves and intermediate nodes that connect two constituents;
- dependency format allows n-ary connections when several words are connected to one headword; punctilog allows connections only between two elements;
- dependency graph's arcs are labelled by the type of the syntactic relation; punctilog's connections are not named;
- dependency annotation treats punctuation in sentences as tokens the same as words; punctilog ignores all initial punctuation in the sentence.

Due to these differences, the transformation from dependency annotation is not straightforward. Three algorithms have been proposed and developed for the transformation. First of all, the algorithms removed all punctuation from the initial sentence. The next steps consist of successive connections of neighboring words. Each connection forms a constituent that is treated as a new word and can be connected further with another word or constituent.

5.1 Algorithm 1

Each word in the sentence with its headword id is extracted from the conllu format and stored in lists. Then the process of annotation with parenthesis starts. Every two words are taken in the parentheses if the word and its head are the next or the following word. The parentheses are placed repeatedly and the id of the head for the formed constituent is calculated as an average of the heads of the united words. This is done in order to be able to more easily know which constituents are closer to each other.

A while loop is connecting the created constituents to the heads with the closest id number to the dependency averages until only one constituent remains.

5.2 Algorithm 2

After removing punctuation, all words with hyphens before or after are connected with corresponding neighboring words.

Then, connections are performed in two cycles: one from the first word to the last one and another in reverse order: from the last word to the first one. For each word connection is performed if its head is a

neighboring word; the same is applied to the already formed constituents. All formed constituents keep all id of their components and id of their heads. The process of connection stops when only one constituent remains.

5.3 Algorithm 3

This algorithm also connects all words to their neighboring heads to form the constituents. The difference is that the formed constituents obtain the id of the head and its connection; the id of the dependent element is lost. This is why only leaves (words or constituents that are not heads for other elements) are connected.

The connections are performed in two cycles: The inner one checks and connects, if possible, each word from the last word to the first one and the outer one repeats the inner cycle until at least one connection is performed in the inner cycle and stops when no further connections are possible.

6. Comparison of the Algorithms' Results

We had no "gold corpus" annotated absolutely correctly for the algorithm evaluation; we had to evaluate the algorithm manually. 20 sentences annotated in the universal dependencies convention were selected for the evaluation. After its transformation by all three algorithms, we compared the results. Surprisingly, all three algorithms produced quite different results and after closer examination, we concluded that no one of these results was absolutely correct.

Figure 2 presents the same sentence as in Figure 1, but in the form of a dependency tree created by another conllu viewer[4]. The verb "antrenează" is still a root and all other words are connected to it. In order to facilitate the comparison, an online visualization tool for punctilog format has been developed[5]. Figures 3, 4, 5 present graphical representations of the punctilog format created by the algorithms 1, 2, 3 respectively.

In the figures, it is seen that all algorithms produced different structures. All three connected the predicate (*nu antrenează*) to the subject

[4] https://urd2.let.rug.nl/~kleiweg/conllu/

[5] https://univoc.dev/

(*Un incendiu exterior*) and the complex direct object (*explozia practic instantanee a aproape întregului conținut al ambalajului*) was connected partially. Algorithm 1 connected only the first word of the object (*explozia*) to the predicate and the rest of the object was connected apart. Algorithm 2 also connected only a part of the object (*explozia practic instantanee a aproape*) to the predicate and the rest of the subject was taken apart. Algorithm 3 connected the subject and the predicate and the complex object was all together. This variant was the most similar to human judgment. One problem was that in classical structure the object should be connected to the predicate and only after that the subject and the predicate should be united. The other problem of the result was the connections of the complex object elements.

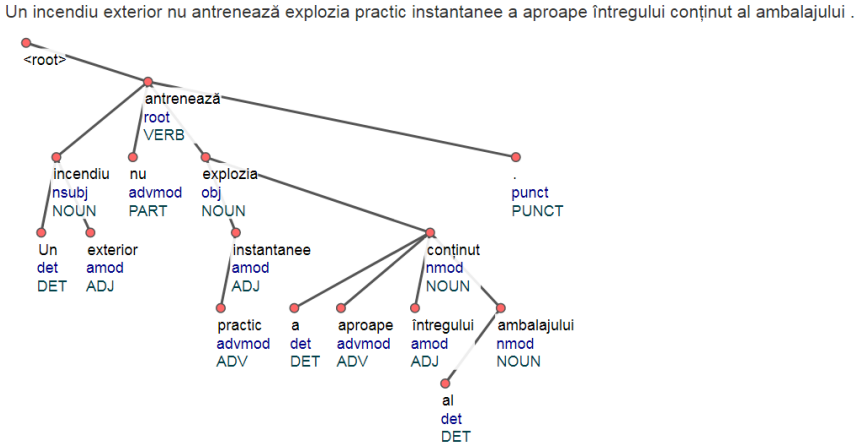

Figure 2. Graphical representation of the universal dependencies annotation of the same sentence as in Figure 1.

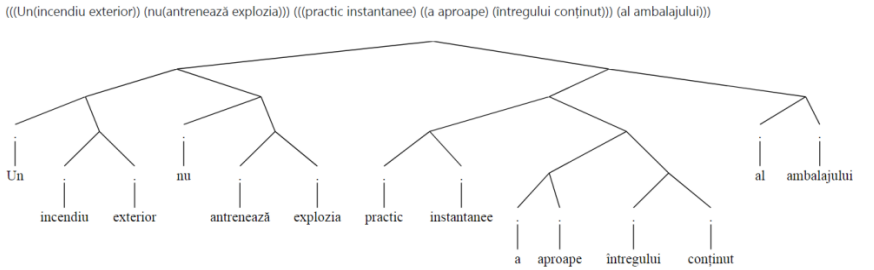

Figure 3. Graphical representation of the punctilog annotation produced by Algorithm 1.

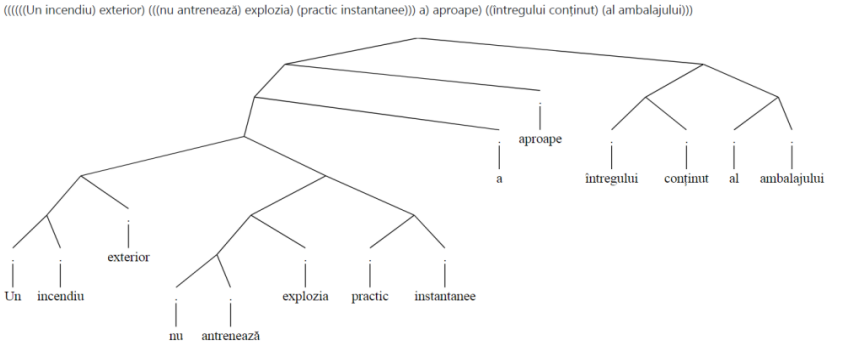

Figure 4. Graphical representation of the punctilog annotation produced by Algorithm 2.

All three algorithms connected this part (*explozia practic instantanee a aproape întregului conținut al ambalajului*) in different ways.

After common discussion between the authors the final and most correct version of the punctilog connections was created; it is presented in Figure 6. The predicate (*nu antrenează*) includes the object (*explozia practic instantanee a aproape întregului conținut al ambalajului*) and all this part is connected to the subject (*Un incendiu exterior*). Complex noun phrase of the object consists of the main part (*explozia practic instantanee*) and the dependent part (*a aproape întregului conținut al ambalajului*) which in turn is formed of parts: (a), (*aproape întregului*) and (*conținut al ambalajului*).

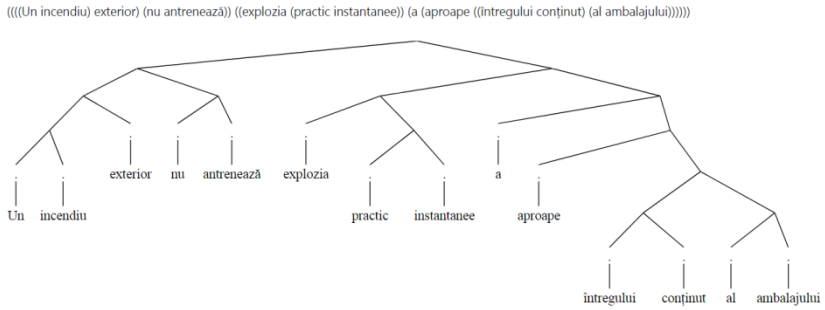

Figure 5. Graphical representation of the punctilog annotation produced by Algorithm 3.

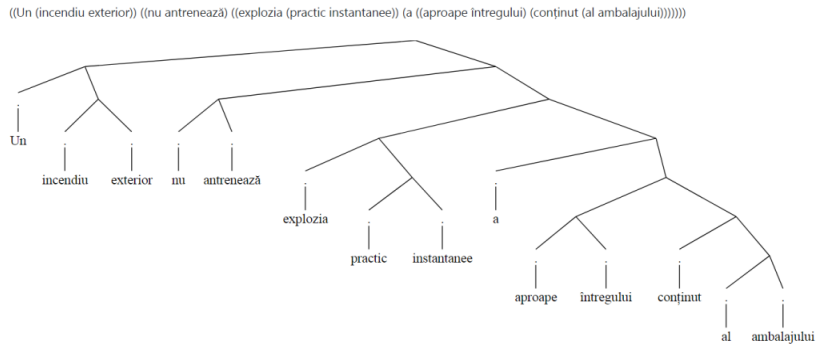

Figure 6. Graphical representation of the punctilog annotation created manually.

7. Discussion and Future Work

As it was discussed in the previous section, we have to modify the algorithms as all of them made different errors in punctilog connections. All algorithms used only the information about the links between words not taking into consideration their morphological and link labels. These labels can be used to connect correctly determiners, adjectives, and articles to the nouns in noun groups; verbs with their auxiliary verbs, adverbs, and particles in verb groups and then to connect formed direct object to predicate and finally subject to the rest of the sentence.

Thus, to form correct punctilog constituents, we have to start with the connections inside classical noun and verb phrases [9]. For noun phrases, firstly the modifiers such as adverbs and adjectives are connected to the words they modify; then articles and parts of complex noun phrases. An example is the complex noun phrase: "*explozia practic instantanee a aproape întregului conținut al ambalajului"* (a virtually instantaneous explosion of almost the entire contents of the package). It is seen in Figures 1 and 2 how the words are connected in dependency grammar convention and their parts of speech. Firstly, adverbs are connected to adjectives, then adjectives to nouns, next articles to the formed groups, and finally several nouns with their dependent words are connected together. The order of connections is:

explozia practic instantanee -> explozia (practic instantanee) -> (explozia (practic instantanee));

a aproape întregului conținut al ambalajului -> a (aproape întregului) conținut al ambalajului -> a (aproape întregului) conținut (al ambalajului) -> a (aproape întregului) (conținut (al ambalajului)) -> (a (aproape întregului) (conținut (al ambalajului))).

Finally, these two parts are connected together. However, there is still possible ambiguity. The part *(a (aproape întregului) (conținut (al ambalajului)))* can be connected as in Figure 6 or as *(a ((aproape întregului) conținut) (al ambalajului))*; this version is presented in Figure 7.

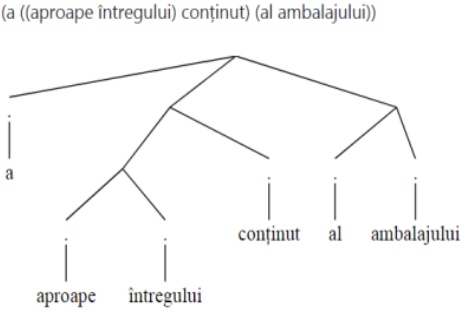

Figure 7. Graphical representation of the alternative punctilog annotation of the complex noun phrase.

The methodology described above in this section is connecting the parts as in Figure 7 as *aproape întregului conținut* is a noun phrase with the main noun and *al ambalajului* is a noun phrase as well. The first word ***a*** with morphological tag *det* and connection to the first noun *conținut* is connected to the first part as well and the structure is slightly different: *((a ((aproape întregului) conținut)) (al ambalajului))*. This structure may also be considered correct.

8. Conclusion

In this paper, ongoing work on the creation of the text corpus annotated with a novel methodology named punctilog is described. The methodology aims to represent the sentence's meaning through its structure. We discuss which structures are appropriate and how to create a corpus of texts annotated in this convention. We present the experiments on the re-annotation of universal dependencies Romanian corpus. We discuss the developed algorithms' results and our future plans for their improvement.

References

[1] N. Chomsky. *The Logical Structure of Linguistic Theory*. Springer, US, 1975, 592 p.

[2] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. *Universal Dependency Annotation for Multilingual Parsing*. In: Proceedings of ACL, 2013.

[3] M.-C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C.D. Manning. *Universal Stanford Dependencies: A cross-linguistic typology*. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, 2014, pp. 4585-4592.

[4] S. Petrov, D. Das, and R. McDonald. *A Universal Part-of-Speech Tagset.* In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012, pp.2089-2096.

[5] Daniel Zeman, et al. *Universal Dependencies 2.8.1*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2021, http://hdl.handle.net/11234/1-3687.

[6] Ioachim Drugus. *PML: A Punctuation Symbolism for Semantic Markup*. In: Proc. of 11th International Conf. "Linguistic Resources and Tools for Processing the Romanian Language", 2015, pp.79-92.

[7] Andrew Carnie. *Constituent Structure*. Oxford University Press; 2nd edition, 2010, 320 pages, ISBN-10: 0199583463.

[8] Jirka Hana and Jan Štěpánek. *Prague Markup Language Framework*. In: Proceedings of the Sixth Linguistic Annotation Workshop, Association for Computational Linguistics, 2012, pp. 12-21.

[9] J. Miller. *A critical introduction to syntax*. London: Continuum, 2011, 275 pages.

[10] Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*, Chicago: University of Chicago Press, 1994.

[11] K. Vijay-Shanker and Aravind Joshi. *Unification-Based Tree Adjoining Grammars*, 1991. Technical Reports (CIS).

[12] Daniel Sleator and Davy Temperley. *Parsing English with a Link Grammar*. In: Third International Workshop on Parsing Technologies, 1993.

Ioachim Drugus[1], Tudor Bumbu[1, 3], Victoria Bobicev[2], Victor Didic[2, 3], Alina Burduja[2], Alexandr Petrachi[2], Victoria Alexei[2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E-mail: ioachim.drugus@math.md, tudor.bumbu@math.md

[2]Technical University of Moldova
E-mail: victoria.bobicev@ia.utm.md, alina.burduja@iis.utm.md, alexandr.petrachi@iis.utm.md, victoria.lazu@ia.utm.md, victor.didic@iis.utm.md

[3]Est Computer
E-mail: bumbutudor10@gmail.com, victor.didic864@gmail.com

Advanced pre-hospital triage based on vital signs in mass casualty situations

Constantin Gaindric, Sergiu Șandru, Sergiu Puiu, Olga Popcova,
Iulian Secrieru, Elena Guțuleac

Abstract

In case of disasters, the pre-hospital triage is a very important stage, crucial for providing the necessary medical assistance and facilitating casualty evacuation from the disaster site to the nearest specialized hospitals. When mass casualty situations take place, casualty prioritizing for every triage category (Red, Yellow, Green) could help to improve the distribution of the available resources (healthcare personnel, ambulances) and, finally, save more lives. This article describes how a computer-aided tool for pre-hospital triage based on vital signs can be used in practice on the disaster site, as well as its possible applications in training purposes.

Keywords: medical informatics, pre-hospital triage, mass casualty situations, knowledge acquisition, computer-aided tool.

1. Introduction

The efficiency of medical first aid, especially in mass casualty situations, is extremely dependent upon the time of provided treatment. In the disaster area, usually, the number of casualties could be extremely high, many of them require urgent medical assistance simultaneously, but the available capabilities and resources (such as healthcare personnel/aides and ambulances) are limited.

Therefore *pre-hospital triage* is one of the most important elements in crisis management response in large-scale disasters. Pre-hospital triage helps to classify rapidly casualties in various homogeneous categories, taking into account the severity of injuries, in order to provide efficient

medical assistance and evacuation from the impact zone to the nearest specialized hospitals.

To support the pre-hospital triage process, as a part of the decision support framework for the management of complex mass casualty situations [1], we are developing a computer-aided tool for pre-hospital triage. This tool is aimed to gather primary medical data of casualties, to assess the triage category via rule-based decision support, and to give the possibility of *setting-up a priority within every triage category*.

2. Methodology

Commonly the following 3 basic categories are used for casualties assessment in the triage methods and algorithms:

- Red (Absolute emergency) – Life-threatening casualties with serious and very serious injuries, illnesses, intoxication, or contamination, who require immediate stabilization measures, as well as priority evacuation in assisted medical transport conditions.
- Yellow (Relative emergency) – Casualties with serious or moderate injuries, illnesses, intoxication, or contamination, with retained vital functions, but with the risk of developing life-threatening complications immediately ahead. They require urgent medical assistance, but not an immediate one.
- Green (Minimal emergency) – Casualties with minor injuries, illnesses, intoxication or contamination, no life-threatening, which can be treated later, usually in outpatient conditions. They can be evacuated in non-specialized transport or independently.

Some triage approaches consider additional categories such as Orange [2] or Gray [3] in order to use more effectively the medical personnel. This becomes especially important under the resources scarcity or when the road infrastructure also was affected by disaster, making casualties transportation problematic.

The need for casualty prioritizing in every triage category (Red, Yellow, Green) in pre-hospital conditions seems to be an advanced step, helping to improve the distribution of the available resources and, finally, save more lives.

3. Designing a computer-aided tool for pre-hospital triage

We have studied and analyzed different clinical and emergency guides and protocols, including the national ones – for Moldova [4]-[5]. As a result, we have selected the following basic attributes (casualty characteristics) which determine the decisions for triage based on vital signs, given in Table 1, and allow quick categorization of casualties:

Table 1. Basic attributes and values in triage based on vital signs

	Red (I)	Red (II)	Yellow	Green
Glasgow Coma Scale	3-8	9-13	14	15
Airways Permeability	Obstruction / Stridor	Difficult breathing	Normal breathing	Normal breathing
Pulse	>120 or <40	111-120 or 41-45	81-110 or 46-59	60-80
Systolic Blood Pressure	<80	80-89	90-100	>100
Respiratory Rate	>35 or <13	29-35	19-28	14-18
Oxygen saturation	<= 85	86-90	91-95	96-100
Individual mobility	Unable	Unable	With help	Walking

The Glasgow Coma Scale is used to objectively evaluate a person's level of consciousness after an injury. Its assessment is based on three aspects of responsiveness: eye-opening, motor, and verbal responses.

All these attributes and values allow us to create decision rules to distinguish priority I and priority II in the Red triage category, supporting the *decision-making*. In addition, our computer-aided tool will provide the possibility to end-user (as an option) to set up a priority within every triage category, helping to follow the casualty more accurate status, avoiding under-triage and over-triage, and suggesting life-saving interventions for casualty, needed in every specific case (with some prioritizing in the chain of emergency care).

4. Conclusions and future work

The standard clinical protocol in emergency cases, designed mainly for a single or limited number of persons, is not suitable for direct use on-site in case of large-scale disasters. Also limitation of the number of triage groups only to three (Red, Yellow, and Green) can lead to some problems in case of mass casualty situations, having a lot of persons in Red and Yellow categories in a short period of time.

Therefore there is a need to advance the algorithm and reasoning for casualty prioritizing for every triage category (Red, Yellow, Green), based not only on theoretical knowledge but on practical experience.

The described approach for advanced pre-hospital triage based on vital signs in mass casualty situations represents the basis for the elaboration of different applications having multiple purposes:

- Being developed as a computer-aided tool on mobile devices, it can be used by the rescue teams members in practice on disaster sites.
- Being implemented as a web application, it can be used for both teaching and training of paramedics, and for their evaluation (determining the level of knowledge and practical skills).

Also, this application can be used to interact with experts for the acquisition of new knowledge (by analyzing non-ordinary cases and collecting data on possible different opinions for specific cases) or for validation of the detected cut-off thresholds, which can be used to stratify casualties.

If it is possible, prioritizing in the Yellow category is extremely desirable, of course, in case the needed resources are available.

The validation process will include obtaining explanations and interpretations of all conclusions, including intermediate ones, obtained in the process of using the proposed pre-hospital triage.

Acknowledgments. The Moldovan State Program project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data” and the project G5700 “An Adaptive Decision Support Framework for the Management of Mass Casualty via an Artificial Intelligence Based Multilayered Approach integrating an Intelligent Reachback Information System” in the

framework of the NATO Science for Peace and Security Programme have supported part of the research for this paper.

References

[1] C. Gaindric, S. Cojocaru, S. Pickl, S. Nistor, Iu. Secrieru, O. Popcova, D. Bein, and D. Cimpoesu. *A Concept for a Decision Support Framework for the Management of Complex Mass Casualty Situations at Distribution Points.* In: Proceedings of the Conference on Mathematical Foundations of Informatics MFOI'2018, July 2-6, 2018, Chisinau, Republic of Moldova, pp. 90-102.

[2] C. Barfod, M.M.P. Lauritzen, J.K. Danker, et al. *Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department – a prospective cohort study.* Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, vol. 20 (2012), article number: 28. DOI 10.1186/1757-7241-20-28.

[3] K. Sakanushi, T. Hieda, T. Shiraishi, et al. *Electronic triage system for continuously monitoring casualties at disaster scenes.* Journal of Ambient Intelligence and Humanized Computing, 4 (2013), pp. 547-558. DOI 10.1007/s12652-012-0130-2.

[4] Gh. Ciobanu, M. Pîsla, F. Gornea, et al. *Ghid naţional privind tirajul medical în incidente soldate cu victime multiple şi dezastre*. Centrul Nat. Şt.-Practic Medicină de Urgenţă, Centrul Republican Medicină Calamităţilor. – Chişinău, 2010, 36 p.

[5] M. Ciocanu, Gh. Ciobanu, V. Cojocaru, A. Oglinda, L. Chiosea, N. Buzatu, and I. Gurov. *Protocol clinic standardizat. Triajul în Unităţile Primiri Urgente*. Chişinău, 2017, 23 p.

Constantin Gaindric[1], Sergiu Şandru[2], Sergiu Puiu[3], Olga Popcova[4], Iulian Secrieru[5], Elena Guţuleac[6]

[1456]Affiliation: Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chisinau
E-mails: constantin.gaindric@math.md, oleapopcova@yahoo.com, iulian.secrieru@math.md, elena.gutuleac@math.md

[2]Affiliation: Emergency Medicine Institute, Chisinau
E-mail: serghei.shandru@gmail.com

[3]Affiliation: Medical Center "AnaMaria-Med", Chisinau
E-mail: puiusv@yahoo.com

Proceedings of Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

On Kyiv Approaches to Knowledge Testing in E-learning

Olena Glazunova, Bella Golub, Vitaly Klimenko, Alexander Lyaletski

Abstract

Kyiv approaches to the construction of systems and tools for the intelligent testing of knowledge obtained by a trainee in remote learning using e-learning systems and/or e-textbooks are considered. Analytical and deductive types of intelligent testing distinguished from the question-answer testing method are described. It is demonstrated that there exist necessary theoretical and practical results for the construction of intelligent testing systems of new types.

Keywords: intelligent testing of knowledge, question-answer testing, analytical testing, deductive testing, e-learning.

1 Introduction

Currently, there is a large number of "shells" and tools focused on the creation of e-learning systems and/or electronic textbooks for a wide range of disciplines taught in general and higher educational institutions. A characteristic feature of systems of this type is that they are all focused on a very wide area of their application, in connection with which almost all of them are intended for the simplest, "question-and-answer" type of testing the student's knowledge, i.e. when knowledge is tested on the basis of the tutor's indication of the correct answer from a list of options offered by the tutor. That is, the testing process can degenerate into a random choice of answer options. This form of

knowledge testing is not always suitable for physical and mathematical disciplines, assuming that an answer is an analytical (symbolic) expression or a formal proof, that is, a chain of deductive and/or inductive steps that ensure the truth of a statement under consideration.

The current state of informatics in the field of the construction and use of both computer algebra systems and automated reasoning systems has initiated the transition from simple "question-answer" testing to more intelligent types of knowledge testing: analytical and deductive. (The first type is applicable, for example, to testing the solutions of a wide variety of physical and trigonometric tasks both from school and student textbooks. The second can be very useful, for example, in studying various mathematical or other formalizable theories that require deductive constructions.)

Below we give a description of the achievements of Ukrainian researchers in the field of analytical transformations and automated reasoning, which can be used to construct tools for intelligent testing other than the "question-answer" type and requiring the formalization of knowledge in the form perceived by a computer. Due to limitations on the size of a workshop paper, this description takes the form of a list of basic approaches, algorithms, and methods. Additionally, note that in Kyiv all investigations on these research areas were initiated by Academician V.M.Gluskov in the first half of the 1960s.

2 Analytical testing

The need for it arises when a trainee's response must be an analytical (symbolic) expression that is an answer to a problem being solved. To carry out such testing, it is required to have a set of "shells" allowing a computer, using analytical transformations, to make sure that the expression proposed by a trainee is correct, i.e. it can be converted into a formula suggested by a tutor. (Scope: various branches of physics, trigonometry, elementary algebra, transformation of expressions using predefined equations, etc.). For this, first of all, it is necessary to have at least such universal and specialized tools as:

- programs for calculations with unlimited accuracy;

- integer, rational and complex arithmetics;

- universal programs for establishing the equivalence of two symbolic expressions (generally, various rewriting rule systems are developed and use for this);

- methods for transforming mathematical expressions to special mathematical forms;

- tools for formula transformations of mathematical expressions presented in the form of hierarchically given data structures of arbitrary complexity;

- distributed calculations based on the view of data as a collection of separate, related expressions.

There are a lot of computer algebra systems satisfying these requirements (see, for example, the site "https://en.wikipedia.org/wiki/List_of_computer_algebra_systems"), according to which it is reasonable to note that one of them appeared practically before everything others else under the leadership of Academician V.M. Glushkov at the Institute for Problems of Mathematical Machines and Systems of NASU (IPMMS) in the mid-1960s, which leads to the appearance of specialized computers of the MIR line specifically designed for numerical and analytical computations (1965 - 1973). Note that the MIR line computers can be considered as forerunners of the personal ones [1].

The experience gained during the operation of these computers was reflected in the design and implementation of systems belonging to computer algebra systems of the Analitik family [2] (with an input language also named Analitik [3]): "Analitik-79" for the SM 1410 computers (1975 - 1983) as well as "Analitik-89", "Analitik-91", "Analitik-93", "Analitik-2000", and "Analitik-2007" for the IBM PC computers. Working-outs of the "Analitik-2000" system put it, in terms of both functionality and implementation efficiency, on a par with such well-known computer algebra systems as Mathematica, Axiom, Maple, etc. and they even surpass them in some characteristics. That is, the Analitik family systems contain everything that is needed to build various "shells" for efficient analytical transformations, in particular, "shells"

for analytical testing including different symbolic computations such as the efficient establishing of the equivalence of different algebraic and trigonometric expressions and many others.

3 Deductive testing

Deductive testing is based on a deductive paradigm and consists in checking a chain of inferences that can be expressed in some formal language, close to an ordinary language and used in the course of gaining knowledge by trainees. (Areas of possible application: mathematical disciplines that require checking the correctness of the trainee's deductive and inductive constructions; jurisprudence, when the testing of a trainee is to test his ability to draw legally correct conclusions and generate legal acts and/or regulations that do not contradict current legislation; and some others).

The deductive paradigm itself is based on a declarative way of the representing and processing of computer knowledge, when it has a form of formalized texts (for example, in mathematics, in the form of axioms, definitions, theorems, etc.) and the testing of a trainee's knowledge consists in checking the correctness of building a chain of logical steps leading to a desired result. Such knowledge processing systems are called automated reasoning systems, most of which represent the so-called automated theorem proving systems, since namely the logical-mathematical approach turns out to be the most relevant and effective one both in an automated search for logical inference and in the verification of a formal (not necessarily mathematical) text on the correctness of its conclusions having the form of an obvious deduction (from the point of view of a computer).

To solve the problems of deductive testing, we adhere to a number of natural requirements for systems of this kind:

- the language for the writing of reasoning conducted by a trainee should be a formal one close to languages of natural publications, thereby allowing to preserve the structure of a problem being solved and to translate it into a form adapted to its processing on a computer.

- each step of deduction, given in a tested text, must be "understandable" by a computer in the sense of the possibility of verifying the text's correctness.

- along with the universal methods for logical inference search, there must be a (refill) set of heuristic inference methods, including inductive deduction methods.

- accumulated knowledge (i.e., learned by students) should be stored as a hierarchical (ontological) information environment and should be used in the usual way for regular training courses; they should be constantly updated by the data adsorbed by a trainee.

In Ukraine, the beginning of a serious study of the deductive paradigm was made by V.M. Glushkov in the paper [4], in which he proposed the basic principles of building automated reasoning systems. As a result, by 1978, at the Glushkov Institute of Cybernetics, a Russian-language automated theorem proving system was designed and implemented. It partially reflected the requirements listed above for deductive testing. Much later, in 1998, the work began on the creation of an English-language version of this system, which took the name of System for Automated Deduction, SAD, currently available online on the website "nevidal.org" (additional information on SAD can be found in [5, 6]). Now, the research on SAD is moved to the National University of Life and Environmental Sciences of Ukraine (NULES).

The SAD system has a three-level architecture and includes the following (see, for example, [7] and [8]):

- an English version of a formal natural language being close to the usual language of mathematical publications;

- a translator from this language into some kind of first-order language called ForTheL [9] in order to provide the ability to carry out a proof search in first-order logic and to make the verification of mathematical texts [10];

- a module for making deduction and verification in syntactic units of this language, reflecting generally accepted (heuristic) methods of reasoning such as dividing problems into subtasks, simplifying problems, some types of (mathematical) induction;

- an efficient method for a sequent-type inference search, which can carry out deduction in the signature of an original theory and use, if necessary, tools for analytical transformations mentioned above and powerful, external w.r.t. SAD, automated theorem proving systems (provers), such as SPASS, Vampire, and Otter.

As interesting experiments carried out with SAD, we can mention the verification of the finite and infinite versions of Ramsey's theorem, some properties of finite groups, the irrationality of the square root of a prime number, Tarski's fixed point theorem, and some others.

The above-said demonstrates that the SAD system can be taken as a prototype for the development of systems, methods, and tools for the deductive testing of a trainee knowledge.

4 Conclusion

The analysis of the state-of-art of the intelligent information technologies in Ukraine shows the following: researchers representing IPMMS and NULES have everything necessary to pass to qualitatively new, distinguished from the question-and-answer type, intelligent types of the testing of the knowledge of a trainee, being trained in a certain e-learning course. This can lead to a more objective assessment of a trainee knowledge and, as a result, to the in-depth and careful studying of disciplines admitting at least a partial formalization and being of the form of special courses and electronic textbooks.

The next, obvious step is the integration and cooperation of analytical and deductive testing tools, which gives a transition to an even more intelligent form of knowledge testing.

Having the above-described implemented/adapted methods and tools for intelligent testing, the improvement of the quality of e-learning can be achieved first, by incorporating these methods and tools into already existent learning system taking into account peculiarities of subject area under consideration and second, by carrying out a full cycle of designing an e-learning course (textbook) based on intelligent testing, that is by developing an (original) electronic "shell" for a created course

(textbook), filling it with intelligent test tools and equipping it with all question-and-answer capabilities that are necessary for objectively checking the quality of a trainee knowledge.

Additionally note that using an approach from [11], it is possible to pass to multi-language intelligent testing.

References

[1] V.P. Klimenko and Yu.S Fishman. *The first personal computers in the world.* Cybernetics, no 3, 1993, pp. 176–180. (In Russian).

[2] A.A. Morozov, V.P. Klimenko, and A.L. Lyakhov (Eds). *Computer algebra systems of Analotik family. Theory. Realization. Application.* Manuscript, K., 2010, 764 pp. (In Russian).

[3] V.M. Glushkov, V.G. Bodnarchuk, T.A. Grinchenko, A.A. Dorodnitsyna, V.P. Klimenko, A.A. Letichevskii, S.B. Pogrebinskii, A.A. Stognii, and Yu.S Fishman. *Analitik (an algorithmic language for the description of computing processes using analytical transformations).* Cybernetics, no 3, 1971, pp. 102–134. (In Russian).

[4] V.M. Glushkov. *Some problems in the theories of automata and artificial intelligence.* Cybernetics, no. 2, 1970, pp. 3–13. (In Russian).

[5] A. Lyaletski, M. Morokhovets, and A. Paskevich. *Kyiv School of Automated Theorem Proving: a Historical Chronicle.* In book: "Logic in Central and Eastern Europe: History, Science, and Discourse", University Press of America, USA, 2012, pp. 431–469.

[6] A. Lyaletski. *Evidence Algorithm and SAD Systems: Past and Possible Future.* Cybernetics and System Analysis, vol. 57, 2021, pp. 9–16.

[7] A. Lyaletski, A. Lyaletsky, and A. Paskevich. *Evidential paradigm and SAD systems: features and peculiarities.* International Journal of Mathematical Sciences and Computing, No. 2, 2018, pp. 1–11.

[8] A. Lyaletski and A. Lyaletsky. *Evidence Algorithm approach to automated theorem proving and SAD systems (In honor of 50 years of Evidence Algorithm announcement).* Selected papers of the 7th International conference "Information Technology and Interactions (IT&I-2020)", Kyiv, Ukraine, December 02-03, 2020, CEUR Workshop Proceedings, vol. 2833, 2021, pp. 144–163.

[9] K. Vershinin and A. Paskevich. *ForTheL – the language of formal theories.* International Journal of Information Theories and Applications, Volume 7, Issue 3, 2000, pp. 120–126.

[10] A. Lyaletski, A.Paskevich, and K. Verchinine. *SAD as a mathematical assistant – how should we go from here to there.* Journal of Applied Logic, vol. 4, no. 4, 2006, pp. 560–591.

[11] E. Glazunova, B. Golub, and A. Lyaletski. *On a multi-language computer support of a human mathematical activity.* In: Proceedings of the 15th International Conference on ICT in Education, Research and Industrial Applications: Integration, Harmonization and Knowledge Transfer (ICTERI 2019), Kherson, Ukraine, June 12-15, 2019, pp. 98–101.

O. Glazunova[1], B. Golub[1], V. Klimenko[2], A. Lyaletski[1]

[1] National University of Life and Environmental Sciences of Ukraine
E–mail: o-glazunova@nubip.edu.ua, bellalg@it.nubip.edu.ua, a.lyaletski@nubip.edu.ua

[2] Institute for Problems of Mathematical Machines and Systems of NASU
E–mail: klimenko@immsp.kiev.ua

Proceedings of the Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

Artificial Intelligence in Dentistry: Teeth Classification

Corina-Elena Iftinca, Adrian Iftene, Lucia-Georgiana Coca

Abstract

Artificial intelligence (AI) starts to be used more and more in the medical field, and by extension also in dentistry. Even later AI seems to be embraced with skepticism, in the last years, scientists and dentists united their forces and wrote scripts that can process images, put a diagnostic or recognize teeth. In this paper, we will present a method to recognize teeth from panoramic dental radiographs using Convolutional Neural Networks (CNN). The teeth were classified into eight major classes as follows: Upper Molars, Upper Premolars, Upper Canines, Upper Incisors, Lower Molars, Lower Premolars, Lower Canines, and Lower Incisors.

Keywords: artificial intelligence, convolutional neural networks, image classification.

1. Introduction

This project aimed from the beginning to classify teeth from images using an existing data set or by creating its own set of images. At first, it seemed difficult to make our dataset, as on the Internet we were not able to find anything that could help us. This is due to the fact that in our country the dental offices are private and each dental office has its own database with panoramic dental radiographs. Second, because of the GDPR law from May 2018 on the protection of persons concerning the processing of personal data. On the panoramic dental radiography appeared the name and the date of birth of the patient. Also in the countries where scripts used in dentistry were written the data scientists collaborated with university hospitals with large databases of panoramic dental radiographs, so it was not difficult for them to make their dataset.

Remains still the question: *how have we managed to make our dataset*? Most of the panoramic dental radiographs came from two doctors from Iasi. To respect the law, before processing the panoramic dental radiographs for obtaining the teeth, we have first cropped the personal data's from the radiographies.

Another reason for choosing this theme for this project is because we think that recognizing the teeth on panoramic dental radiographs is the first step in putting a diagnostic on the teeth, as long as before knowing what problem a tooth has we have to localize it. We hope that this project will be a start point for another project with a more complex approach, such as writing scripts that could put a certain diagnostic on a tooth.

2. State of the Art

In [1], it is shown the fact that "*AI (including ML) has already invaded and established itself in our daily lives*". The purpose of AI is to complete tasks that are currently done by people with greater accuracy, speed, and lower resources, but for this, it is necessary to increase the computer's power. AI is perfect for work that implies a large amount of data that needs to be processed. Also, AI is better for repetitive activities due to the lack of fatigue and its higher performance. AI and ML started to be used more and more in the field of diagnostic imaging in dento-maxillo-facial radiology [1], but also in other domains from medicine [2], [3] like in dialysis and kidney transplant [4], cancer [5], tuberculosis [6], [7], and stroke [8].

The future directions for using AI in dental radiology are to diagnose osteoporosis, identification of periapical disease, or classification the tumors. The AI software needs to be trained on huge datasets in order to recognize the patterns. The accuracy for a certain diagnostic depends on the algorithms used, the labels or the images used that should be representative ones. AI software should be able to understand new information, make "*intelligent decisions*" and "*learn from mistakes to improve the decision-making*" for future images. Unfortunately, until now, the AI applications were not feasible to be used on a large scale in routine dentistry, although the progress is exponential.

In [9], it is specified the fact that "*digitalization in dentistry has increased significantly over the last 10-20 years*" and that there is a need

for AI software especially in the countries where there is a lack of specialists. The paper presents the AI approaches for different fields of dentistry. In orthodontics, for example, the AI was used to predict which tooth should be removed in the orthodontic treatment as this decision relies on the dentist's experience. In conservative dentistry and prosthodontics, the AI was used to "*determine the most suitable material for the restoration of cavities and long-term monitoring of the reconstruction process*". In periodontology, temporomandibular joint disorders and maxillofacial surgery AI technologies were used in order to diagnose a certain disease as for endodontics AI was used to establish the working length of the root canals.

In [10], faster regions were combined with convolutional neural networks (faster R-CNN) in order to detect and recognize the teeth in dental periapical films. For improving detection precision, three post-processing techniques were used: (1) a filtering algorithm that deleted overlapping boxes detected by faster R-CNN associated with the same tooth, (2) a neural network that detected missing teeth, and (3) a rule-base module for numbering teeth used for detect results that break certain intuitive rules. The precision and the recall were calculated for the test dataset, and their values were over 90%. The test dataset was also manually annotated by three other doctors and the conclusion was that the algorithm results were similar with a junior dentist level.

In [11], convolutional neural networks were used for tooth detection and recognition in panoramic radiographs. For training the network, a number of 1352 panoramic radiographs of adults were used. The teeth detection module was used to define the boundaries of each tooth. Like in the previous paper, R-CNN architecture was involved. The teeth numbering module classified the teeth detected in the images according to the FDI notation. For the test set, a number of 222 panoramic radiographs were used. For the teeth detection, the system achieved a 0.9941 value for sensitivity and 0.9945 for precision, while the experts detected teeth with a precision of 0.9998 and a sensitivity of 0.998. Regarding teeth numbering, the system achieved a value of 0.98 for sensitivity and 0.9994 for specificity while the experts had 0.9893 for sensitivity and 0.9997 for specificity. The conclusion was that the performance of the algorithm was similar to the expert one.

3. Dataset

In Figure 1 shown below, it is presented the flowchart for obtaining the dataset, a short preview for the details given in the next subsections.

Figure 1. Flowchart for a dataset.

The dataset for the experiments was obtained from panoramic dental radiographs. "Panoramic radiographs consist of a series of narrow tomograms sequentially scanned onto the detector (film or storage phosphor in a cassette, or a solid-state digital detector) beneath a secondary slit. Panoramic radiology aims to produce a complete view of both dental arches and their adjacent structures with minimal geometric distortion and with minimal overlap of anatomic details from the contralateral side" [12]. In Figure 2, a panoramic radiograph is shown.

Figure 2. Panoramic radiograph.

On both dental arches there are 32 teeth, identified by a unique two-digit combination. The first digit specifies one of the four quadrants of the mouth as it follows: the maxillary right quadrant has assigned the number 1, the maxillary left quadrant has assigned the number 2, the mandibular left quadrant has assigned the number 3, and the mandibular right quadrant has assigned the number 4. The second digit indicates the tooth within the quadrant. In every quadrant, the teeth are numbered, mesial to

distal, from 1 through 8, beginning with the middle incisor and ending with the third molar [13].

Because usually, the third molar is not present on the dental arches it wasn't taken into consideration. When the creation of the dataset was started, in the notation of the images was used the FDI two-digit system, and a number in parenthesis which represented the number of the X-ray as shown in Figure 3. After cropping all the teeth from the panoramic dental radiographs (Figure 4), eliminating the ones with cavities, fillings, or dental crowns it was concluded that the experiments would have better results if instead of 28 classes it would be used 8 bigger classes, as it can be seen in Table 1, assigned in 8 folders as shown in Figure 5.

Table 1. The 8 classes used to classify our dataset in order to make the experiments

Upper teeth	**Lower teeth**
• Upper molars • Upper premolars • Upper canines • Upper incisors	• Lower molars • Lower premolars • Lower canines • Lower incisors

The teeth were cropped manually, because they didn't have the same position on different X-rays and this is why it was impossible to write a script that would have done the crops automatically.

Figure 3. The notation first used in naming the cropped images. In the above figure, there are some images from the Lower canines' folder.

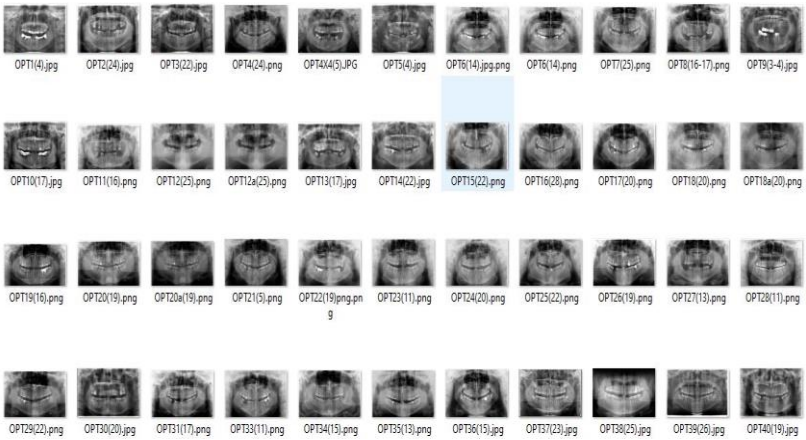

Figure 4. The panoramic dental radiographs from which the teeth were cropped in order to obtain the dataset classified in the 8 classes. OPT comes from orthopantomography which is another term for panoramic dental radiography.

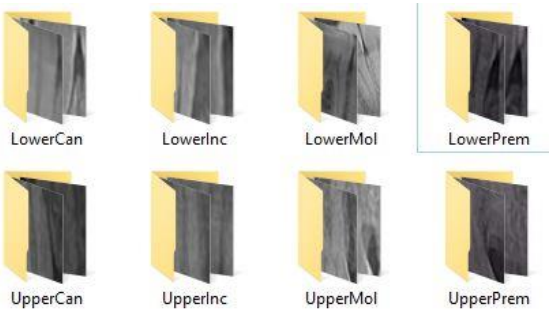

Figure 5. The folders with the cropped images, classified into 8 main classes that will be used for making the experiments.

After obtaining a total of 772 teeth, by using scripts, the images from the dataset were resized. This process was made by keeping the original image and adding a black or a white background around it to obtain the same size for all the images. In Figure 6, it is shown an example.

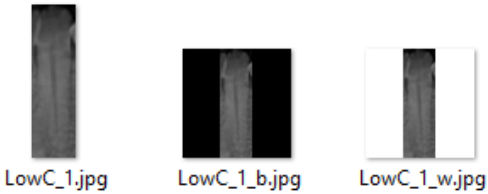

Figure 6. From left to right: the original crop of a lower canine, the resized image with black background, the resized image with white background.

There were made six folders, as shown in Figure 7:

- three folders for each size: 32 × 32, 64 × 64, and 128 × 128 pixels with white background;
- three folders for each size: 32 × 32, 64 × 64, and 128 × 128 pixels with black background.

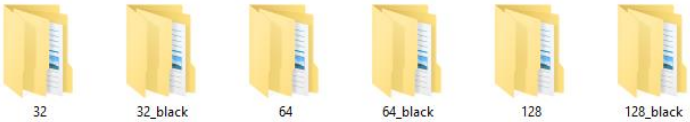

Figure 7. The six folders with the datasets.

Each of the six folders contained another 3 folders:

- Train;
- Validation;
- Test.

4. Dataset Evaluation

4.1 Building the teeth classification model

In order to build the model, Google Colab was used, as it had already all the libraries needed for the experiments and the possibility to connect with the personal Google drive account where the dataset was imported. As the images cannot be fed directly into the model, a NumPy array of images was created. Also, all the images were normalized meaning that the value of the pixels, which were between 0 and 255, were divided by 255 in order to be in the [0, 1] range.

The model was defined using Keras similar to models from [14], [15]. A CNN was used for this work. As an optimizer, the one that updates the weight parameters to minimize the loss function, we have chosen Adam. For the loss function, the one that tells the optimizer if it is moving in the right way in order to achieve the global minimum, we have chosen sparse categorical cross-entropy. The model was trained for 50 epochs.

4.2. Experiments

The experiments were made on three different image sizes: 32 × 32, 64 × 64, and 128 × 128 pixels. For each size, the model was trained using 2 backgrounds: white and black.

The dataset was split as follows:

- train: 80%;
- validation: 10%;
- test: 10%.

After running all the experiments, for the best and the worst results, a data augmentation was made, in order to see how much this influences the results. Also, because the dataset was unbalanced, for the test data another four metrics besides accuracy were used: F1, precision, recall, and confusion matrix.

4.3 Results

Regarding the images with the size of 32 × 32 pixels, in Figure 8, we can see the plots for accuracy and loss for train and validation datasets for the images with black background.

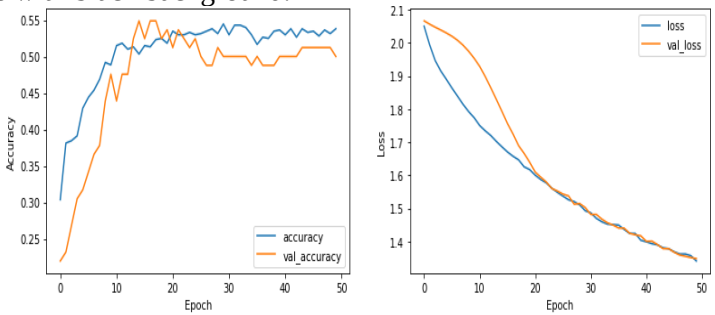

Figure 8. Plots with accuracy and loss for train and validation datasets for images with the size of 32 × 32 pixels.

For the images with the size of 64 × 64 pixels, in Figure 9, we can see the plots for accuracy and loss for the train and validation datasets.

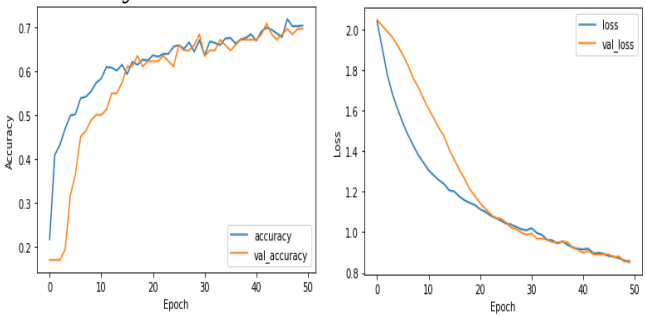

Figure 9. Plots with accuracy and loss for the train and validation datasets for images with the size of 64 × 64 pixels.

For the images with the size of 128 × 128 pixels, in Figure 10, we can see the plots for accuracy and loss for the train and validation datasets.

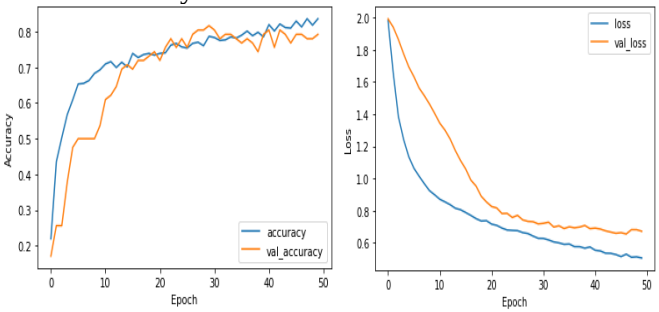

Figure 10. Plots with accuracy and loss for the train and validation datasets for images with the size of 128 × 128 pixels.

The best results, as we can see from the plots and the tables above, were obtained for the images with the size of 128 × 128 pixels, the same as in the case of the images with white background.

Data augmentation was used for the datasets where we have obtained the best and the worst results. We are referring to the images with the size of 128 × 128 pixels with black background for the best result and to the images with the size of 32 × 32 pixels with black background for the other case. Regarding the images with the size of 32 × 32 pixels, in

Figure 11, we can see the plots for accuracy and loss for the train and validation datasets. Also, in Table 2, there are shown the results for accuracy and loss for the model we have trained with and without data augmentation.

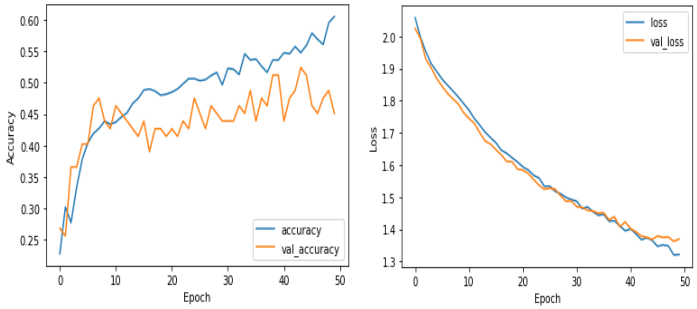

Figure 11. Plots with accuracy and loss for the train and validation datasets for images with the size of 32 x 32 pixels after data augmentation.

Table 2. Values for loss and accuracy for images with the size of 32 x 32 pixels.

	Train	**Validation**	**Test**
Accuracy	0.544	0.548	0.573
Loss	1.340	1.347	1.279
Accuracy after data augmentation	0.605	0.524	0.630
Loss after data augmentation	1.320	1.362	1.268

For the dataset with images having the size of 128 × 128 pixels, in Figure 12, we can see the plots for accuracy and loss for the train and validation dataset. Also, in Table 3, there are shown the results for accuracy and loss for the model we have trained with and without data augmentation.

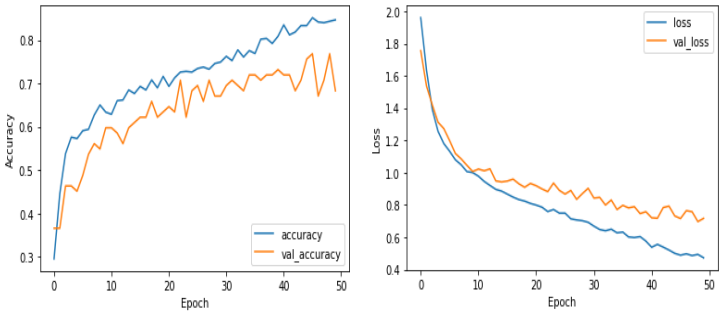

Figure 12. Plots with accuracy and loss for the train and validation dataset for images with the size of 128 x 128 pixels after data augmentation.

Table 3. Values for loss and accuracy for images with the size of 32 × 32 pixels.

	Train	**Validation**	**Test**
Accuracy	0.836	0.817	0.750
Loss	0.504	0.654	0.749
Accuracy after data augmentation	0.851	0.768	0.761
Loss after data augmentation	0.472	0.696	0.731

As we can observe, from the tables and the plots above, there is a slight improvement in accuracy, for the train dataset and test dataset. Better results are obtained after data augmentation for the smaller images, while for the other one, the difference between the results with and without data augmentation is smaller.

5. Conclusion

Teeth classification is a challenge as:

- on the Internet there are no datasets available with panoramic radiographs so the first desideration before starting this work, was getting the dataset involved in the experiments;

- data refining must be done carefully because teeth do not have the same size;
- it requires a correct annotation in order to correctly classify the teeth.

Although having a large dataset with images is an important thing when we speak about classification, smaller datasets such as the ones used in this work, can obtain good results when a model is trained on them.

Acknowledgments. This work was support by project REVERT (taRgeted thErapy for adVanced colorEctal canceR paTients), Grant Agreement number: 848098, H2020-SC1-BHC-2018-2020/ H2020-SC1-2019-Two-Stage-RTD.

References

[1] T. Joda, M. Bornstein, R. Jung, M. Ferrari, T. Waltimo, and N. Zitzmann. *Recent trends and future direction of dental research in the digital era*. International Journal of Environmental Research and Public Research, (2019).

[2] A. Iftene. *Using Artificial Intelligence in Medicine*. In: Proceedings of the Conference on Mathematical Foundations of Informatics MFOI2020, January 12-16, 2021, Kyiv, Ukraine, (2021), pp. 161-169.

[3] A. Burlacu, A. Iftene, E. Bușoiu, D. Cogean, and A. Covic. *Challenging the supremacy of evidence-based medicine through artificial intelligence: the time has come for a change of paradigms*. In: Nephrology Dialysis Transplantation (ndt), gfz203, Oxford Academic, (2019), pp. 1-4.

[4] A. Burlacu, A. Iftene, D. Jugrin, I.V. Popa, P.M. Lupu, C. Vlad, and A. Covic. *Using Artificial Intelligence Resources in Dialysis and Kidney Transplant Patients: A Literature Review*. BioMed Research International, 2020, article ID 9867872, (2020), 14 pages.

[5] A.J. Banegas-Luna, J. Pena-Garcia, A. Iftene, F. Guadagni, P. Ferroni, N. Scarpato, F.M. Zanzotto, A. Bueno-Crespo, and H. Perez-Sanchez. *Towards the Interpretability of Machine Learning Predictions for Medical Applications Targeting Personalised Therapies: A Cancer Case Survey*. International Journal of Molecular Sciences, vol. 22, issue 9, 4394, (2021).

[6] A. Hanganu, C. Simionescu, L.G. Coca, and A. Iftene. *UAIC2021: Lung Analysis for Tuberculosis Classification*. In: Proceedings of the Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum.

Bucharest, Romania, September 21-24, 2021, vol. 2936, (2021) pp. 1253-1263. http://ceur-ws.org/.

[7] L. G. Coca, A. Hanganu, C.G. Cușmuliuc, and A. Iftene. *UAIC2020: Lung Analysis for Tuberculosis Detection*. In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020. vol. 2696, (2020) http://ceur-ws.org/.

[8] M. Rezmerita, I. Cercel, and A. Iftene. *Stroke Detector - An Application that applies the F.A.S.T. Test to Identify Stroke*. In: Proceedings of the 17th International Conference on Human-Computer Interaction RoCHI 2020, 22-23 October, 2020, (2020), pp. 39-47.

[9] M. E. Machoy, L. Sommerfield, A. Vegh, T. Gedrange, and K. Wozniak. *The ways of using machine learning in dentistry*. Advances in Clinical and Experimental Medicine, (2020).

[10] H. Chen, K. Zhang, P. Lyu, H. Li, L. Zhang, J. Wu, and C. U. Lee. *A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films*. Scientific reports, (2019).

[11] D. Tuzoff, L. Tuzova, M. Bornstein, A. Krasnov, M. Kharchenko, S. Nikolenko, M. Sveshnikov, and G. Bednenko. *Tooth detection and numbering in panoramic radiographs using convolutional neural networks*. Dentomaxillofacial Radiology, (2019).

[12] A. Farman, S. Clark, A. Friedlander, W. Jacobs, Z. Khan, G. Kushner, K. Norman, C. Nortjé, A. Silveira, R. Wood, and S. Yaggy. *Springer Panoramic Radiology* 1st edition, (2007), pp. 7.

[13] J.C. Türp and K. W. Alt. *Designating teeth: The advantages of the FDI's two-digit system*. Quintessence International, (1995).

[14] L.G. Coca, A. Iftene, and T. Manoleasa. *Asphalt crack identification experiments using convolutional networks*. In: International Conference on INnovations in Intelligent SysTems and Applications (IEEE INISTA 2021). Kocaeli, Turkey, August 25-27, (2021).

[15] L.G. Coca, C.G. Cușmuliuc, and A. Iftene. *Automatic tarmac crack identification application*. In: 25rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. 8-10 September. Szczecin, Poland, (2021).

Corina-Elena Iftinca, Adrian Iftene, Lucia-Georgiana Coca

"Alexandru Ioan Cuza" University of Iasi Romania, Faculty of Computer Science
E-mail: {corina.iftinca, adiftene, georgiana.coca}@info.uaic.ro

Proceedings of Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

Generation and use of educational content within adaptive learning

Alexandr Parahonco, Mircea Petic

Abstract

One of the new educational technologies that has shown its undoubted effectiveness is e-learning. Nowadays, beyond e-learning, adaptive platforms have been appeared with commercial services for any sort of adaptation. In the Republic of Moldova where universities are not highly financed, such commercial systems are unprofitable in usage. Our solution lays under development of our own adaptive learning system on the base of the Moodle Web platform.

Keywords: adaptive learning, plug-in, crawler, Flexible, TestWidTheory, TestWid.

1 Introduction

Intelligent adaptive learning systems originate fast, but still incur obstacles in realization. Theoretically these systems must organize learning process based on differentiation for each student. Formative and diagnostic assessment should apply adaptive intelligence for precise results. The problems arise in the ways of implementation of these principles and involved technologies.

Thus, in the 21st century, the idea of adaptive learning becomes even more popular: not only teachers and psychologists, but also managers and businessmen are interested in it. Subsequently, such popular adaptive educational systems as Knewton Alta, Smart Sparrow, Geekie, ALEKS and others appeared. Such systems to a greater extent

are highly specialized (to the domain of studies, for example) and paid. That is why higher education institutions from Republic of Moldova cannot use them. Our universities need free platform without any domain limitation.

The purpose of the article is to study the principles of generating educational content and the development of an adaptive learning system for higher education institutions of Republic of Moldova that will be capable to compete against leading commercial products.

The article begins with an analysis of the systems of adaptive learning, the development of their common models of adaptive learning and an assessment of their relevance to the usage in the educational process of the university. Then it discusses the developing model for Moodle Web platform as the base for new adaptive learning system. At the end of the article, practical solutions are proposed for the implementation of adaptive learning in the framework of the national system of higher education [1, p. 163].

2 Analysis of the existed systems

Adaptive learning systems represent educational information and communication technologies that respond in real time to student actions and, in accordance with the information received, provide him with individual support [2, p. 8].

When creating an adaptive educational system, first of all, three key questions are solved: what is modeled, how it is modeled and how the adaptation model is supported. Then one of three scenarios is implemented, where the adaptation object can be: **content, tasks or the order of presentation** of educational materials [2, p. 9].

If in the educational system the **object of adaptation** is content, then it functions according to the following algorithm: first, it analyzes the student's response to the task and, in the case of an incorrect answer, offers him feedback, tips, or additional educational materials. For example, Geekie is a paid learning platform powered by artificial intelligence (AI) to prepare Brazilian students for their final exams. This

platform provides adaptation at the level of curriculum modification [3].

Smart Sparrow offers **three levels of customization**: feedback, curriculum modification (learning paths), and the ability of the educator to facilitate the transmission of knowledge. Functionality of this platform allows teachers to create interactive content that can be adapted for any group of students according to the topic of the subject under study and the specific requirements for the group's learning process. Teachers can determine individual learning paths for students, interact with them in real time, and also use a number of ready-made templates to save time when creating electronic content [4, p. 374].

Knewton Alta platform is well known for its programs and applications with adaptive functions. Knewton's team of specialists managed to create universal algorithms and develop an extensive infrastructure for collecting, analyzing and using information about student progress. It contains **two elements of adaptation**: tasks and the procedure for providing materials [5, pp. 202-207].

One of the developers of the adaptive tests for monitoring is NWEA (Northwest Evaluation Association) that creates the adaptive tests for different goals. For example, the test MAP Growth is used for the periodic testing of pupils' knowledge of different subjects, while MAP Skills is recommended to be applied more often [6, pp.553-554].

All these educational platforms – Knewton Alta, Geekie, Smart Sparrow and NWEA – are not suitable for the university's higher education system, because they are expensive, overly flexible, which is inconsistent with the structure and time constraints of courses; are time consuming.

To summarize, we propose to create new adaptive learning system on the base of Moodle Web platform. Moodle competes on an equal footing with the world flagships of the distance education system market. Thanks to this, Moodle combines a wealth of functionality, flexibility, reliability and ease of use. That is why, most of our universities use this educational platform what played an important role in our decision.

3 Moodle model of adaptive learning. Adaptive test

Due to the fact that it is a "plug-in base system", we developed at first model for adaptive learning. To form an adaptive course, we offer a model (Fig. 1), which consists of three separate plugins: a plugin for adaptive testing **TestWid**, a plugin for storing educational materials (text, video, audio, images, exercises, tests) **TestWidTheory**, and a plugin for an adaptive course **Flexible**. The degree of adaptation should cover all three levels: content, order of presentation and tasks.

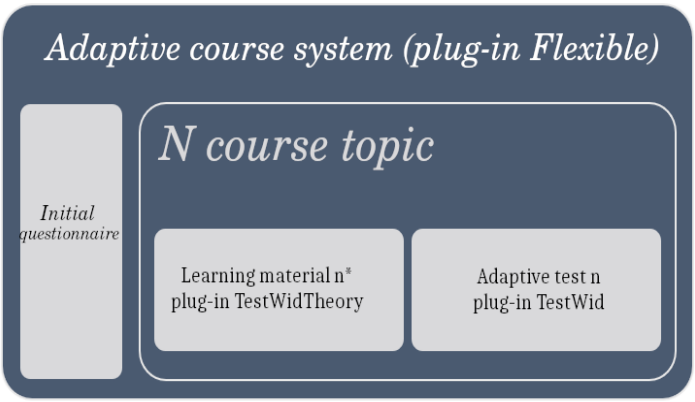

Figure 1. Adaptive course model for the Moodle platform.

This system should allow the teacher to generate course content and supplement it for each student, based on his preferences. These preferences will be determined in the **initial survey** (questionnaire) with 2 simple questions with the list of answers: 1) sort in descending order types of learning materials according to your opinion (videos, images, text, learning games); 2) sort in descending order the source list for custom content search (google, youtube, wikipedia, encyclopedias, others).

When saving this data, the system (Flexible plug-in) will arrange

learning resources in each topic by priority.

The developing system requires that the assessment of the quality of students' knowledge and their competencies is determined using an adaptive test (TestWid) for each topic. It is not prohibited to use other plugins (test, assignment and others) to control the quality of training, but the assessment will not be considered the main one. This rule also applies to plugins for learning resources: at least one instance of the TestWidTheory plugin must be in each theme. The idea of using these two plugins together is dictated by the connection of each piece of theory with a specific question from the test. So, a teacher primarily creates the content, then selects by the mouse some fragments and, clicking the button "Adaugă teoria la TestWid", selects from drop-down list from which adaptive test and towards which question to link the fragments.

Thus, if a student has not formed the required set of competencies and has not passed the test, he will be presented with a list of tasks with incorrect answers and a link to related fragments of the theory (Fig. 2, 3).

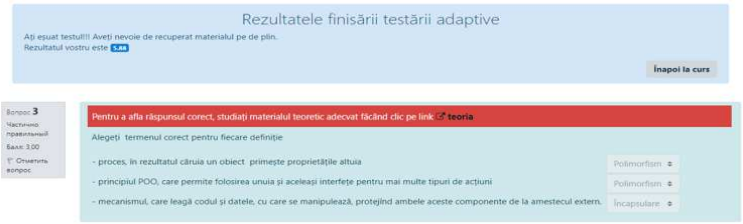

Figure 2. List of incorrectly completed tasks – question number 3.

After repeated self-preparation, the test will be retaken. When retaking, the testing system will select questions from each category that are different from those already used, if any.

The development of an adaptive testing model for the TestWid plugin was based on the **"Three-level algorithm"** model [7, pp. 233-234], which takes into account the capabilities of the Moodle platform

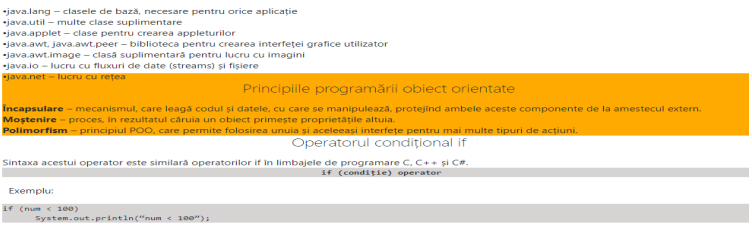

Figure 3. Fragment of theory related to question 3 after following the link.

(fixed number of questions in the test). The "three-level algorithm" allows, in the presence of 15 questions-tasks, to achieve the same accuracy and reliability as in the test with 45 exercises that do not pay attention to their level of complexity, and also allows three times to reduce the cost of testing duration, while maintaining information security.

Therefore, the question with the **adaptive path** of the student is solved. Against this background the question arises about both answer assessment and formula of the final grade. The solution comes from the **mathematical model** for assessing knowledge based on **learning levels**. The characteristic of an assignment is the level of assimilation, for which it is intended to test. Tasks can be divided into five groups corresponding to the levels of assimilation: understanding, identification, reproduction, application, creative activity [8, p. 12-13]. A set of essential operations is determined for each task. Essential operations are those operations that are performed at a verified level.

Thus, to assess students' answers and knowledge, the coefficient K_a is used (1):

$$K_a = \frac{P_1}{P_2}, \tag{1}$$

where P_1 – the number of correctly performed essential operations in the control process;

P_2 – the total number of significant operations in the test;

$\alpha = 0, 1, 2, 3, 4$ – corresponds to the levels of assimilation.

The grade is set on the basis of the specified cut-off values by ratios multiplied by 10:

$K_a < 0.7$ – unsatisfactory

$0.7 \leq K_a < 0.8$ – satisfactory

$0.8 \leq K_a < 0.9$ – good

$K_a \geq 0.9$ – excellent.

Finally, due to the "three-level algorithm" and level of assimilation it is possible to create truly adaptive test that can estimate student's knowledge and skills.

4 Model for the dynamic content generation for training courses

According to the object of adaptation, content generation should allow users (teachers and students) to generate learning content. Such kind of system can be developed by the **web-scraping** technology, including **Data mining** and **text mining** approaches. The thorough study of technical documentations, science articles and practical experience, we designed the Scheme of the program model for the dynamic creation of e-courses (see Fig. 4).

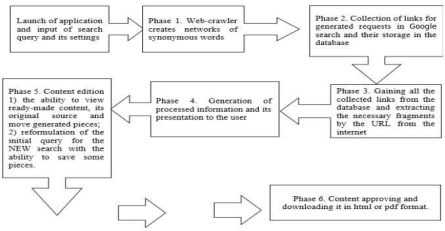

Figure 4. Scheme of the program model for the dynamic creation of e-courses.

As it can be seen, at Phase 1, the operating principle of the developed model is to use synonymous connections to search for dictionaries

that are similar in meaning with the help of a crawler, and use them at Phase 2 and 3 for advanced search using the Google search service. Thus, we obtain the behavior model of a user performing manual scraping.

According to phase 4 and 5, well merged content should be generated and further exported in Moodle. The content may be imported in Moodle via Page and File standard plugins at Phase 6.

Our application will allow editing content (font, placement, color) and downloading the files in html or pdf formats.

5 Description of the design and principle of operation of the Flexible plug-in

All plugins developed within the flexible course were created based on the documentation of the Moodle developers [7] and are designed to work on Moodle platforms starting from version 3.

The work of any plugin on the Moodle platform begins with its creation. To create a flexible adaptive course plug-in, you need to go to the Courses section and select the Add Course option.

Flexible course plug-in was developed from the standard "Topics" course plug-in, as it was consistent with our responsive course model and required fewer programming changes compared to other **formats**: the only element of the course, forum and sections by week.

The structure of the Flexible course is provided by an algorithm that consists of **two phases**: creation and updating. In the first phase, it checks the number of plugins in each section of the course and, if they are not there, adds instances of the TestWidTheory and TestWid plugins to each section using the **completeStructure** function. In the second phase, which occurs when the user adds new sections, our algorithm runs the already mentioned completeStructure function to create the described course structure. Thus, the plug-in mechanism creates a layout for the Flexible course.

6 Description of the design and operation of the TestWidTheory plug-in

As mentioned earlier, the adaptive course plugin consists of two plug-ins: TestWidTheory and TestWid. The plugin for storing educational materials (TestWidTheory) (see Fig. 5) is the prototype of the standard Moodle LMS "Page" plug-in.

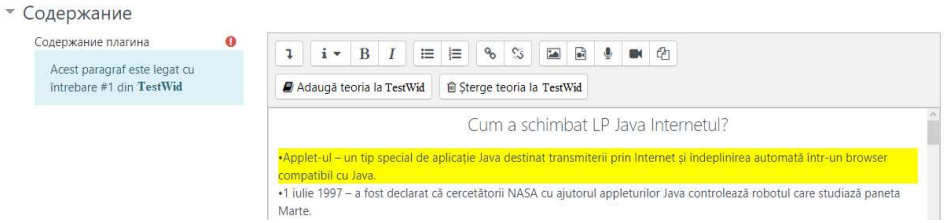

Figure 5. Highlighting a related piece of theory.

This plug-in is a Web-based WYSIWYG editor that is needed to store learning resources (text, video, audio, images) and create links to specific questions from the adaptive test in the current section (see Fig. 2, 3). The work with the plugin for storing educational materials includes the following steps:

1. highlight a fragment of the theory;
2. click on the button "Adaugă teoria la TestWid";
3. in the opened modal window, select the required adaptive test, the question number and click on the "Salvează" button ;
4. Save changes to the plugin.

It should be noted that steps 1 - 3 must be completed 15 times according to the number of questions in the adaptive test. For comfortable work with the plug-in, all related fragments are highlighted when the mouse cursor is hovering.

7 Description of the recovery process and retake

The recovery process is necessary for students with a coefficient $K_a < 0.7$ (grade less than 7), which indicates the incompetence of the students. In this case, after the end of the test, they go to the section with the results. This section can also be accessed through the section "Assessments" (Grades).

In this section, this category of students has the opportunity to study all questions with **partially correct** or **completely incorrect** answers. As it is seen from Fig. 2, each question is accompanied by a link to the relevant theoretical course material. At the same time, the student does not see his previous answers, which makes the recovery process transparent.

After preparation, students take the same adaptive test again. The "TestWid" plugin uses the **load_not_used_questions** function to determine the previously used questions and generate new questions of the corresponding levels. This procedure lasts until either student finally takes the test with satisfactory or bigger grade, or the number of unresolved questions is run out.

8 Testing the developed plug-ins

The study used testing both to create the necessary functions, methods, systems, and to check the quality of their work in general. The final check was carried out on the Moodle Web platform version 3.5.1+ of Alecu Russo Balti State University, and the local one is on version 3.9.1+. Testing was carried out at the following **levels**: unit testing; integrating testing; system testing.

Within each level of testing, the following testing **methods** were used based on manual testing: Installation testing, Usability Testing, Volume Testing, White box, Black box, Grey box, and Graphical user Interface Testing.

The final testing was held on 02.04.2021 at the Alecu Russo

Balti State University, in groups Mathematics and Computer Science (MI21Z) and Computer science (exact sciences) (IS21Z) with overall 26 students, united into one experimental group while studying the course "Programarea orientată pe obiect II (Programarea Java)".

After passing the adaptive test, students passed a questionnaire to assess its work. 18 out of 26 students took part in the survey (69.23%). 55.56% of them agreed that our adaptive testing is better than traditional testing and 44.44% have the opposite opinion.

9 Conclusion

One of the new educational technologies that has shown its undoubted effectiveness is e-learning. After analyzing the principle of operation of the Knewton Alta, Geekie, Smart Sparrow, and NWEA adaptive learning systems, we came to the conclusion to create our adaptive learning system on the base of Moodle Web platform, consisting of three plug-ins: Flexible, TestWid and TestWidTheory.

At the moment, a model of adaptive learning on the Moodle platform has been implemented partially. Only TestWid and TestWidTheory plug-ins functionality has been developed and tested. In the future, our research provides for the full implementation of this model.

Acknowledgments. This article was written within the framework of the research project "20.80009.5007.22 Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

References

[1] A. Parahonco. *Exploring the capabilities of adaptive learning systems.* In: The Technical Scientific Conference of Undergraduate, Master and PhD Students, vol 1, 2020, pp. 163–166, ISBN 978-9975-45-632-6.

[2] K.A. Vilkova and D.V. Lebedev. *Adaptive learning in higher education: pros and cons.* Modern Education Analytics, vol. 7, no. 37, p.9, ISSN 2500-0608.

[3] *How software that learns as it teaches is upgrading Brazilian education.* [online]. [cited 03.01. 2020]. Available from: https://www.theguardian.com/technology/2016/jan/10/geekieeducational-softwarebrazil-machine-learning.

[4] V. Bocharov and L. Suslova. *Adaptive learning in non-formal education.* In: VI International Scientific and Technical Conference "Modern Information Technologies in Education and Scientific Research" (SITONI-2019), 2019, pp.371–376.

[5] D.A. Bogdanova. *About an adaptive platform for individual training.* In: XI International Scientific and Methodological Conference "New educational technologies in Higher Education", 2014, pp. 202–207.

[6] K. Osadcha, V. Osadchyi, S. Semerikov, H. Chemerys, and A. Chorna. *The Review of the Adaptive Learning Systems for the Formation of Individual Educational Trajectory.* CEUR-WS, 2020, pp. 547–558.

[7] Fundamentals of psychodiagnostics. *Textbook for students of pedagogical universities / under the general editorship of A. G. Shmelev.*, Phoenix Publishing House, 1996, 544 p.

[8] A.V. Solovov, A. Menshikova, and L.Klentak. *Methodological foundations of the design of electronic educational resources: a textbook.* Samar Publishing House, 2013, 180 p., ISBN 978-5-7883-0931-6.

Alexandr Parahonco[1], Mircea Petic[2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science, Alecu Russo Balti State University
E–mail: alexandr.parahonco@usarb.md

[2]Vladimir Andrunachievici Institute of Mathematics and Computer Science, Alecu Russo Balti State University
E–mail: mircea.petic@math.md

Proceedings of Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

GeomSpace, an Interactive Geometry Software for Arbitrary Dimensional Euclidean and non–Euclidean Spaces

Alexandru Popa

Abstract

This paper describes GeomSpace, an interactive geometry software. This application is not limited in space dimension or to Euclidean space. It can operate equally well also for Riemannian, Bolyai–Lobachevsky, Galilei, Minkowski, De Sitter, Anti de Sitter, and other homogeneous spaces.

Some theoretical background for GeomSpace is also given.

Keywords: homogeneous space, interactive geometry software, GeomSpace.

1 Motivation

In the last decades, homogeneous spaces are widely studied [2, 3]. However, this study reaches some difficulties. There is no common terminology among different spaces. Based on practical needs of concrete study, different terminology and assumptions are used. Additionally, there is a large number of different homogeneous spaces. The majority of non–linear homogeneous spaces were never described or studied.

In the domain of interactive geometry software, almost all applications work with 2D or 3D Euclidean space. Remaining non–Euclidean projects are usually focused on elliptic or hyperbolic spaces.

GeomSpace [1] is an interactive geometry application for Euclidean and non–Euclidean spaces — Fig.1. Some of its features include:

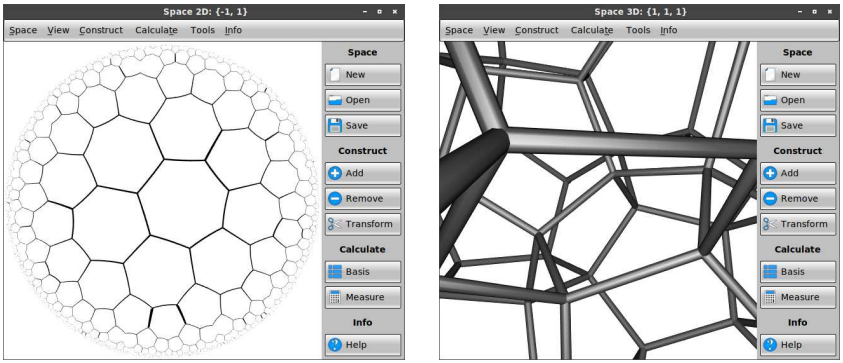

Figure 1. Hyperbolic plane (left) and elliptic space (right) presented in GeomSpace.

- Has no restrictions on space dimension.
- Performs equally well for spaces: Euclidean, Riemannian, Bolyai–Lobachevsky, Galilei (including curved ones), Minkowski (including curved ones De Sitter and Anti de Sitter), and other homogeneous spaces. GeomSpace has no routines specific to some space.
- Gives the possibility to interact with space and figures in it, without worrying about space model.
- Shapes of points and lines are native for space geometry.
- Illumination uses optics laws native for space geometry.
- Uses native OpenGL calls and takes advantage of GPU 3D capabilities not only for Euclidean–like motions, but for all possible motions.

GeomSpace permits to:

- Create different spaces,

- Construct geometric figures in space,
- Vizualize constructed figures and navigate in space with mouse and keyboard,
- Edit constructed figures by adding, removing or moving its elements,
- Perform different calculations: find coordinates, basis vectors, distances, and angles,
- Exchange created figures in own GeomSpace format: binary *.gmsp* or text *.gms*.

2 Notions and definitions

Any geometric quantity (angle, length, area, volume), besides the *value*, also has some *quality*: elliptic, parabolic, or hyperbolic.

Definition 1 [4]. *The characteristic k of a geometric quantity is the number: $k = 1$ for elliptic quality, $k = 0$ for parabolic (linear), and $k = -1$ for hyperbolic.*

Homogeneous space geometry is completely determined by characteristics of its main measurements: lengths between points on lines (k_1), angles between intersecting lines on planes (k_2), dihedral angles between 2-planes in 3-subspace which intersect in line (k_3),..., dihedral angles between hyperplanes in space which intersect in (n-2)-subspace (k_n).

Definition 2 [4]. *The specification of some homogeneous space is the list of its main characteristics: $\{k_1, k_2, \ldots, k_n\}$. The specification of n-dimensional homogeneous space contains n characteristics.*

For some homogeneous space $\mathbb{B}^n = \{k_1, \ldots, k_n\}$, consider linear $(n+1)$-dimensional vector space $\mathbb{R}^{n+1}$. Consider indexed dot product as:

$$x \odot_i y = \sum_{j=0}^{n} K_{ji} x_j y_j, \quad K_{pq} = \begin{cases} 1, & p = q, \\ \prod_{l=p+1}^{q} k_l, & p < q, \\ K_{qp}, & p > q. \end{cases} \tag{1}$$

Definition 3 [4]. *Square matrix $(n+1)\times(n+1)$ is named generalized orthogonal with respect to specification $\{k_1,\ldots,k_n\}$ if all its column vectors $m_i, m_j, i \geq j = \overline{0,n}$ satisfy the condition:*

$$m_i \odot_i m_j = \delta_{ij}. \tag{2}$$

Definition 4 [2]. *Define generalized cosine, sine and tangent as functions $C(x)$, $S(x)$ and $T(x)$:*

$$C(x) = C(x(k)) = \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n}}{(2n)!} = \begin{cases} \cos x, & k = 1, \\ 1, & k = 0, \\ \cosh x, & k = -1; \end{cases} \tag{3}$$

$$S(x) = S(x(k)) = \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n+1}}{(2n+1)!} = \begin{cases} \sin x, & k = 1, \\ x, & k = 0, \\ \sinh x, & k = -1; \end{cases} \tag{4}$$

$$T(x) = T(x(k)) = \frac{S(x(k))}{C(x(k))} = \begin{cases} \tan x, & k = 1, \\ x, & k = 0, \\ \tanh x, & k = -1. \end{cases} \tag{5}$$

In these equalities, the parameter k is the characteristic of argument x, which is always some geometric quantity.

3 Homogeneous space model and its representation in OpenGL

It is common practice in 2D and 3D graphics software to represent displacement of figures with one additional space dimension. This additional dimension is also capable of holding the information about space shape, which is useful in case of not linear spaces.

Lemma 1. *[4] Given a generalized orthogonal matrix M of specification $\{k_1,\ldots,k_n\}$, the following equality holds:*

$$x \odot_i y = Mx \odot_i My \quad \forall x, y \in \mathbb{R}^{n+1}, i = \overline{0,n}. \tag{6}$$

Definition 5 [4]. *Define n-dimensional homogeneous space $\mathbb{B}^n$ with specification $\{k_1, \ldots, k_n\}$ as sphere $\{x \in \mathbb{R}^{n+1} | x \odot_0 x = 1\}$. Define linear space $\mathbb{R}^{n+1}$ as meta–space of homogeneous space $\mathbb{B}^n$. Define origin as vector $e = (1 : 0 : \ldots : 0)$.*

The image of sphere of space is preserved when acting by linear transformations defined by generalized orthogonal matrices.

Definition 6 [4]. *Define space motion (isometry) $\mathfrak{M}$ as linear transformation defined by the respective generalized orthogonal matrix M restricted to $\mathbb{B}^n$. Define m-dimensional subspaces $\mathbb{B}^m < \mathbb{B}^n$ as linear span of the first $m + 1$ column vectors of all possible generalized orthogonal matrices M restricted to $\mathbb{B}^n$.*

The linear span of arbitrary collection of vectors restricted to $\mathbb{B}^n$ need not to be congruent with any subspace of the same dimension. Such linear figures are named *lineals*. All subspaces are also lineals.

Lemma 2. *[4] The distance $d(X, Y)$ between any two points $X, Y \in \mathbb{B}^n$ satisfies the equality:*

$$C(d(X, Y)) = X \odot Y. \tag{7}$$

This distance is invariant with respect to space motions.

GeomSpace makes use of indexed dot product, generalized orthogonal matrices and generalized trigonometric functions to represent homogeneous spaces, interact with their elements and compute their measurements.

OpenGL graphic library uses homogeneous coordinates x, y, z, w, which correspond to linear meta–space $\mathbb{R}^4$. The transformation between OpenGL coordinates (x, y, z, w) and screen space coordinates (x', y', z') is given for linear representation (Beltrami–Klein) by:

$$(x', y', z') = \left(\frac{x}{w}, \frac{y}{w}, \frac{z}{w}\right) \tag{8}$$

and for conformal representation (Poincare disk) by:

$$(x', y', z') = \left(\frac{x}{1+w}, \frac{y}{1+w}, \frac{z}{1+w}\right). \tag{9}$$

4 Example of usage

Consider the problem:

Given: Two lines a and b on $\mathbb{E}^2$ (Euclidean plane).

To find:

1. Determine if the lines a and b are intersected or parallel.
2. (a) If a and b are intersected, find the angle between them.
   (b) If a and b are parallel, find the distance between them.

Usually, in order to resolve this problem, it is necessary three different algorithms:

1. Algorithm of detection the lines parallelism.
2. Algorithm of computing the angle between intersecting lines (the distance between them is irrelevant in this case).
3. Algorithm of computing the distance between parallel lines (the angle between them is irrelevant in this case).

Now, let us change the problem:

Given: Two lines a and b on $\mathbb{S}^2$ (elliptic plane).

To find: The angle between a and b.

Since all lines intersect each other on elliptic plane, the problem greatly simplifies. Nevertheless, the algorithm of computation of angle between intersecting lines cannot be taken from the previous example. We need the completely new algorithm for elliptic plane.

Change the problem again:

Given: Two lines a and b on $\mathbb{H}^2$ (hyperbolic plane).

To find:

1. Determine if the lines a and b are intersected, parallel or divergent.
2. (a) If a and b are intersected, find the angle between them.
   (b) If a and b are parallel, find the inclination between them.
   (c) If a and b are divergent, find the distance between them.

Not only the new relative position of lines appears (divergent), but all algorithms, including the algorithm of relative position determination, need to be revised for hyperbolic plane. As earlier, each algorithm is only applicable in appropriate case.

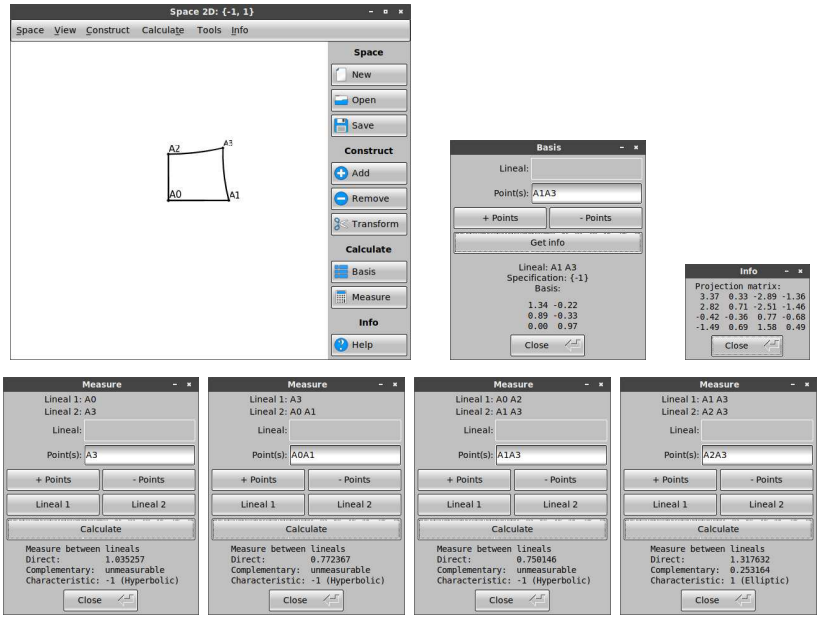

Figure 2. Different measurements on hyperbolic plane. First row: quadrilateral being inspected, line basis, generalized orthogonal matrix of projection. Second row: distance between points, distance between a point and a line, distance between lines, angle between lines.

GeomSpace uses one generic algorithm: computation of measurement between lineals. This algorithm works correctly in all cases and additionally determines relative position of lineals (see Fig.2). The algorithm works in all homogeneous spaces. So, the problem can be formulated as follows:

Given: Two lineals a and b on a homogeneous plane $\mathbb{B}^2$.

To find:

1. Determine relative position between a and b.
2. Find the most appropriate measurement between a and b.

The following information is obtained as the result:

- Characteristic of measurement (elliptic, parabolic or hyperbolic),
- Value of measurement.

Euclidean plane has specification $\{0, 1\}$, so the possible measurement characteristics are:

- 0 (parabolic) when the lines are parallel, or
- 1 (elliptic) when the lines are intersected.

Elliptic plane has specification $\{1, 1\}$, which leaves place for only one measurement characteristic: 1 (elliptic). On elliptic plane this characteristic can be either of angle or of distance. But on elliptic plane there is no distinction between them.

Hyperbolic plane has specification $\{-1, 1\}$, therefore the possible characteristics are:

- 1 (elliptic) when the lines are intersected,
- 0 (parabolic) when the lines are parallel, and
- -1 (hyperbolic) when the lines are divergent.

The algorithm needs not to know a priori what relative positions between two lineals are possible on certain homogeneous plane. It works correctly for all of them. The lineal specification together with characteristic and value of measurement between them gives all necessary information about the relative position of lineals.

It is important, because there exist much more complex situations. For example, on De Sitter plane with specification $\{-1, -1\}$, there are ten possible relative positions of lineals (use term "line" when the lineal is a line, and "lineal" when it is not a line):

1. Two intersecting elliptic lineals with hyperbolic angle between them;

2. Elliptic lineal intersecting parabolic limit lineal with infinite hyperbolic angle between them;

3. Elliptic lineal intersecting hyperbolic line with hyperbolic complementary angle between them (direct angle is not measurable in this case);

4. Two intersecting parabolic limit lineals with infinite hyperbolic angle between them;

5. Two parallel parabolic limit lineals with parabolic distance between them;

6. Parabolic limit lineal intersecting hyperbolic line with infinite hyperbolic angle between them;

7. Parabolic limit lineal parallel to hyperbolic line with parabolic inclination between them;

8. Two intersecting hyperbolic lines with hyperbolic angle between them;

9. Two parallel hyperbolic lines with parabolic inclination between them;

10. Two divergent hyperbolic lines with elliptic distance between them.

5 Conclusion

GeomSpace offers an unusual approach to interactive geometry software, where space geometry is taken as configuration parameter. This approach has advantages and disadvantages. On the one hand, GeomSpace gives less primitives of inspection or interaction with figures compared with other software. For example, there is no notion of parallelism in GeomSpace, because this notion is specific to linear spaces. On the other hand, GeomSpace provides the bird view of different spaces. The same experience of inspection or interaction with figures is possible in different geometries. Such experience gives the feeling of different spaces, their peculiarities and differences between them.

References

[1] *GeomSpace project.* http://sourceforge.net/projects/geomspace/.

[2] B.A. Rosenfeld. *Non-Euclidean geometries.* (in Russian) GITTL, Moscow, 1955, 744 p.

[3] I.M. Yaglom. *A Simple non-Euclidean Geometry and its Physical Basis.* Springer–Verlag, New York, 1979, 307 p.

[4] A. Popa. *Analytic Geometry of Homogeneous Spaces.* arXiv:1807.10134 [math.HO], 2018, 190 p.

Alexandru Popa[1]

[1]SSI Schaefer
E–mail: alpopa@gmail.com

Proceedings of Workshop on Intelligent Information Systems
WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova

Collaborative learning modelled by High-Level Petri nets

Inga Titchiev

Abstract

The rapid development of internet technology induces a new style of learning, which is different from the traditional one and comes to complete it with new opportunities. The study described in this article the researched management of the learning progress and collaborative issues in distance learning by means of Petri nets.

Keywords: e-learning, collaborative learning, Petri nets, efficiency of learning.

1 Introduction

Increasing the efficiency of the educational system, extension and diversifying the educational offers, continuous training by exploitation the opportunities offered by information and communication technologies are the development priorities of the educational system in the Republic of Moldova.

Until recently, in many countries, distance learning has not been widely used for several objective reasons – mainly due to the insufficient development of technical means of training. Currently, the technical premises for the widespread use of distance learning in education have been created, and the COVID-19 pandemic has energized and accentuated their urgent need.

It is essential to identify the needs of users and to integrate into the system the functionalities that allow them to be satisfied. Even if

the emphasis shifts from teacher to student, it is still necessary to obtain information about student progress and the level of collaboration with other students, so it is proposed to use Petri nets [4] as a tool in modeling the management of these processes.

2 Distance education

Distance education [6] gives to learner flexibility in time and location. Thus, being geographically dispersed, they have opportunities to collaborate and to develop even in crisis situations.

Definition 1 [7]. *Collaborative learning is the educational approach of using groups to enhance learning through working together. Groups of two or more learners work together to solve problems, complete tasks, or learn new concepts.*

Collaborative learning involves new opportunities for learners and new approaches for teachers, their role being no less necessary, as a mentor, thereby optimizing the teaching process by distributing the resources of the trained and organizing activities through new technologies.

In order to increase the efficiency of group learning, information and communication technologies in education are coming. An important role in increasing the effectiveness and efficiency of collaborative learning [5] is the motivation of each member of the group, the number of members, their skills.

The proposed approach of modeling the individual and group route management through Petri nets, the MAETIC learning method (from the French: Pedagogical Method with ICT tools), will be applied, which is based on project-based development.

3 High-Level Petri Nets

High Level Petri-nets (HLPNs) asset:

1. High Level Petri-nets have an intuitive graphical representation

and a well-defined semantics that unambiguously define the behaviour of each HLPNs.

2. They are very general and can be used to describe a large variety of different systems.

3. HLPNs have a very few, but powerful, primitives, an explicit description of both states and actions.

4. Are stable towards minor changes of the modelled system.

5. A formal analysis methods allow proving the properties of HLPNs.

6. The two most important analysis methods are known as occurrence graphs and place invariants. Computer tools [8] supporting their drawing, simulation, and formal analysis, exist.

Definition 2 [2]. *A* High-level Petri Nets *is a structure* $HLPN = (P; T; D; Type; Pre; Post; M_0)$*, where*

- P is a finite set of elements called Places.
- T is a finite set of elements called Transitions disjoint from $P(P \bigcap T = \emptyset;)$.
- D is a non-empty finite set of non-empty domains, where each element of D is called a *type*.
- $Type : P \bigcup T \rightarrow D$ is a function used to assign types to places and to determine transition modes.
- $Pre; Post : TRANS \rightarrow \mu PLACE$ are the pre and post mappings with

  $TRANS = \{(t; m) | t \in T; m \in Type(t)\};$

  $PLACE = \{(p; g) | p \in P; g \in Type(p)\}.$
- $M_0 \in \mu PLACE$ is a multiset called the initial marking of the net.

A *Marking* of the HLPN is a multiset, $M \in \mu PLACE$.

A transition is enabled with respect to a *net marking* or in a particular *transition mode*. A transition mode is an assignment of values to the transition's variables, that satisfies the transition condition (i.e., the transition condition is true). The transition's variables are all those variables that occur in the expressions associated with the transition. These are the transition condition and the annotations of arcs involving the transition.

A finite multiset of transition modes, $T \in \mu TRANS$, is enabled at a marking M iff $Pre(T_\mu) \leq M$.

A step may occur resulting in a new marking M' given by $M' = M - Pre(T_\mu) + Post(T_\mu)$.

3.1 Mapping High-Level Petri Nets for collaborative learning

In this section, we use HLPNs [3] to construct various sequence control in distance learning. Depending on the behavior of the learner we can have different learning paths. In order to identify these paths, we will specify several control sequences that may occur. Based on the same course content, we can have different instructional strategies: linear, choice, and arbitrary traces (which combines the first two).

For linear learning path, the learner progress is in a pre-determined order (Figure 1). In Figure 1, the learner can go to the second topic only after finishing the first one, and so on.

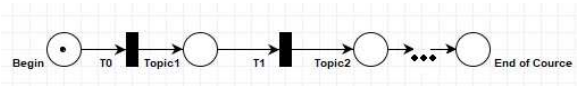

Figure 1. Liniar learning path by Petri nets

Linear choice path allows jumping and selecting the next content in the arbitrary order (Figure 2). The learner in Figure 2, can access any topic he wants from n existing ones arbitrarily.

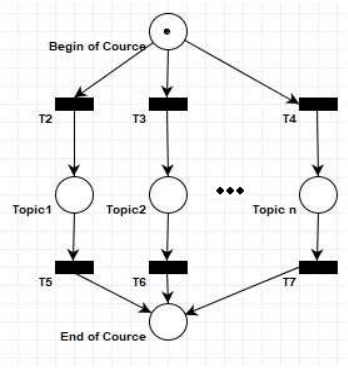

Figure 2. Choice learning path by Petri nets

The collaborative learning is an effective method in distance education, thus we propose to model this process for better understanding.

The goal of collaborative learning is to form a group with heterogeneity even if they have different backgrounds, various learning paths and diverse instruction styles. In order to achieve this goal, it is necessary that each learner has the opportunity to make a break point (jump) to obtain additional information from the outside (a sub-net), so that on return he/she can ensure the homogeneity of the group. After the modeling of collaboration learning trace, by analysing the HPNs, we can estimate the block, deadlock of the system, learning path in order to improve the process.

4 Conclusion

In this article it was shown that Petri nets is a convenient formal method for modeling the management of the learning progress and collaborative issues. Thus, it was shown how to identify the needs of users and integrate the functionalities of the system for better understanding of these processes.

Acknowledgments. 20.80009.5007.22, "Intelligent Information

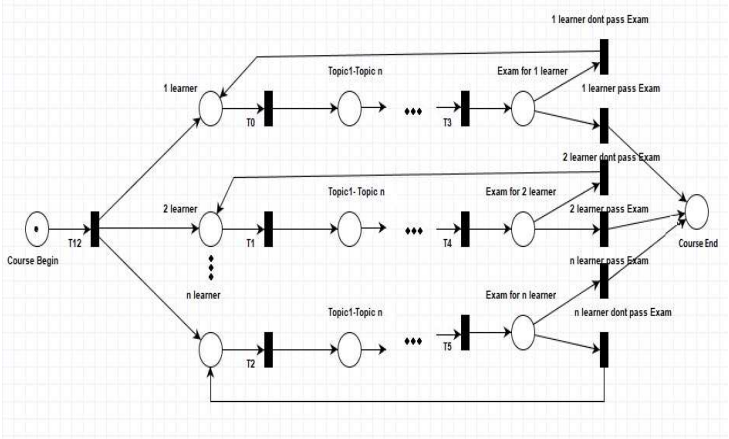

Figure 3. Collaboration modeled by H-L Petri nets

systems for solving ill-structured problems, knowledge and Big Data processing" project has supported part of the research for this paper.

References

[1] X. He, and T. Murata. *High-Level Petri Nets – Extensions, Analysis, and Applications*. Electrical Engineering Handbook (ed. Wai-Kai Chen), Elsevier Academic Press, 2005, pp.459–476.

[2] K. Jensen. *An Introduction to High-level Petri Nets*. In: Proceedings of the 1985 International Symposium on Circuits and Systems: Kyoto 85, pp 723–726, Kyoto, Japan, 1985.

[3] K. Jensen, and G. Rozenberg. *High-level Petri Nets: Theory and Applications*. Springer-Verlag, 724 p., London, UK, 1991.

[4] S. Cojocaru, M. Petic, and I. Titchiev. *Adapting Tools for Text Monitoring and for Scenario Analysis Related to the Field of Social Disasters*. In: Proceedings of The 18th International Conference

on Computer Science and Electrical Engineering (ICCSEE 2016), October 6-7, 2016, Prague, Czech Republic, pp. 886–892.

[5] E. Stacey. *Collaborative Learning in an Online Environment*. International Journal of E-Learning & Distance Education, vol. 14, no. 2, pp. 14–33.

[6] H.W. Lin, Wen-Chih Chang, George Yee, Timothy K. Shih, Chun-Chia Wang, and Hsuan-Che Yang. *Applying Petri Nets to model Scorm Learning Sequence Specification in Collabrative Learning*. In: Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05), 2005, pp.203-208. doi: 10.1109/AINA.2005.120.

[7] *Collaborative learning*. [on-line] https://www.valamis.com/hub/collaborative-learning.

[8] *HiPS: Hierarchical Petri net Simulator*. [on-line] http://hips-tools.sourceforge.net/wiki/index.php/Main_Page.

Inga Titchiev[1,2]

[1]Vladimir Andrunahievici Institute of Mathematics and Computer Science

[2]Tiraspol State University
E–mail: inga.titchiev@math.md

A Brief Overview of Kurdish Natural Language Processing

Ashti Afrasyaw JAF, Sema Koç Kayhan

Abstract

Our survey represents an attempt to focus on the obstacles and barriers that researchers face in this field. By classifying the previous works, analyzing the shared similarities and differences among the papers, we aim to guide scientists to the edge of the studies and ease future efforts. The collected 52 research papers (65% of them were published after 2017), is the first attempt to synthesize the work of Kurdish NLP. The general challenges which concern the dialect base and topic-related groupings are the main parts of the analysis. Each work has been associated with its publisher and available resource links.

Keywords: Dialects, Challenges, Kurmanji, Sorani, Kurdish, Natural language processing.

1. Introduction

Natural language processing (NLP) is the field of language study in more than one direction, including language comprehension, speech recognition, syntactic and semantic analysis of sentences. Moreover, grammatical and semantic processing of the language by presenting semantic objects for induction and deduction is part of NLP. For these purposes, computer techniques in both hardware and software are employed [1].

NLP is covered with great effort in so many languages; this expectation is equally true for the Kurdish language. It is the language of more than 30 million people in the Middle East, and other scattered communities around the world [2]. In the last two decades, this specific situation has subtly changed due to the restructuring of the formal federal government of Iraq and recognizing the Kurdish language as the second

formal language across the country besides the Arabic language [3]. Kurdish instantly becomes the formal language of education in local schools and universities in the regional government north of Iraq. But, still, the language is among the less-resourced languages. Moreover, there are other barriers in dealing with the language, no standard Kurdish language for either speaking or writing is available. Although there are two main particular dialects, which are Sorani and Kurmanji, the dialect diversity is of the Kurdish natural language processing challenge. Even these two dialect branches maintain essential differences in written scripts and some grammar features. For instance, the gender distinction is clear in Kurmanji, unlike the Sorani dialect [4].

In our review, besides the technical issues regarding natural language processing and the level achieved in this area, a brief history of the Kurdish language and linguistic points is focused on. The first sections deal with some challenges covered in several works concerning natural processing in the Kurdish language [4, 5].

The second section points out the classification of the studies in the dialect base. The Kurdish language includes dialects (Sorani, Kurmanji, Zazaki, Hawrami, and Kirmashani). The present resources are almost in Sorani and Kurmanji; most of the papers cited cover these two directions [5]. The technical methods and the specific issues solved by published papers represent the third section of our review.

Finally, the collected works have been synthesized and analyzed from a variety of angles, presenting the works' date, efficiency, authors, and place of publication. The detailed chart of these aspects has revealed some possible hidden factors that have enhanced KNLP improvement. Our review considers the first attempt to insight the field of NLP in the Kurdish language.

2. Basic of the Language

Kurdish is considered one of the Indo-European family languages within the Iranian branch [8]. In the direction of counting the number of speakers after Arabic, Persian, and Turkish, it is forth widely using language in the Middle East, within countries like Iran, Turkey, Iraq, Syria, and some other small scattered groups around the world. Its development has faced noticeable external challenges in different directions like geographical,

religious, and economic [9]. In this way, the absence of the standard language and common writing system are the main factors to slow down convey of any language. The dialects of the Kurdish language are five: Sorani (Central Kurdish), Kurmanji (Northern Kurdish), Zazaki, Gorani (Hawrami), and Kirmashani (Southern Kurdish). The Kurmanji Kurdish (Northern Kurdish) is spoken by approximately 14 million people which is the largest community or dialect group of Kurdish [10]. It is widely spoken close to the border of Iraq, Iran, Syria, and Turkey, approximately by 10 million, 2 million, 1-1.5 million, and 1 million, respectively, in each country, and by less than 1 million in Armenia, Azerbaijan, and Georgia [11]. In addition, there is a massive population of Kurmanji speakers in big cities like Istanbul and Damascus. Therefore, about 65% of Kurdish speakers are Kurmanji. It is typically written in a Latin script called Bedirxan.

Most of the Kurdish cultural and written activities started with Kurmanji. For instance, poetry in the 16th century, first Kurdish print and media, first written periodic – in 1898, alphabet book – in 1909, radio broadcast – in 1920, and satellite television – in 1995 [9].

For the Sorani (Central) Kurdish, more than seven million Kurdish speakers use the central dialect. Almost all of these people live in the north of Iraq and northeast of Iran. Sorani is considered the second dialect after Kurmanji [12, 13]. It is an Indo-European branch within the Iranian languages family. It employs Perso-Arabic Alphabets for writing with some modifications.

According to negation, auxiliary usage, valency, and thematic roles, the positions of these markers vary in the Central Kurdish more than in the Northern dialects; concord markers are fixed, following the verb stem [10]. For verbal agreement and alignment, Central Kurdish deploys pronominal enclitics that attach to the direct object for past-tense transitive constructions, whereas the Northern dialects manage ergative constructions [10, 14]. The definite marker -aka is used in the Central Kurdish dialects which is a further point of distinction with the Kurmanji dialect, where this property is absent [2, 14].

Zazaki is another Kurdish language dialect of Eastern Turkey (more commonly known as Zaza, Kirmanjki, Kirdki, and Dimli). Although ethnically the majority of Zaza people are identified as Kurds, it is

typically inconsiderate a part of the Kurdish language group in the narrow sense. Zaza employs Latin script for writing, while it does not perfectly correspond to the Zazaki phonemic inventory. Kurmanji Bedirxan is sometimes chosen by writers to express their linguistic and ethnic solidarity with the Kurmanji [13, 19]. The dialect is very low resourced compared to the previous dialects.

Kirmashani (Southern Kurdish) is spoken primarily in the Khanaqin and Madlin districts of Iraqi Kurdistan and the Kermanshah region of Iran [17]. The Southern Kurdish dialects are also referred to as Pehlewani, Pahlawanikare [2, 17]. The speakers of the Gorani dialect live across Iraq and Iran border in the narrow region [12]. Gorani/Hawrami is predominantly spoken in Kurdistan Province (southwestern corner), Kirmashan Province (northwestern corner), Hawraman region, and Halabja in the north of Iraq. It possesses the ancients literary tradition among Kurdish dialect groups. It was the language of the court of the Ardalan principality in Sanandaj in the 16th century. Yarasani religion is still the belief of many speakers of the dialect, with a considerable number of its religious texts written in Gorani. The most archaic dialect of this group is considered Hawrami or Hawramani. Hawrami intellectuals have been more active in recent years, in the form of holding cultural events and literary conferences, publishing books and periodicals, broadcasting radio and TV programs, and utilizing social media. The medium of education for Gorani/Hawrami speakers is Farsi in Iran and Sorani or Arabic in Iraq [17, 18].

3. Challenges in Kurdish NLP

Any Natural language processing in both directions of NLP processing or NLP generation basically depends on the available data. Data have different forms as text or speech corpora, computational grammars, lexicon, dictionaries, parallel corpora, Worldnet, and treebanks [14]. In the Kurdish language case, lack of standard language for both written and speaking, even the diversities and differences of the inter dialects besides using different alphabets and scripts for writing causing much complexity in building and preparing resources for Kurdish NLP [5]. Apart from the previous facts, some efforts have been done in the direction of building Kurdish language resources for both Kurmanji and Sorani dialects. More

desiccation about this part is mentioned in the Basic Language Resource Building section.

3.1 Dialect and subdialect diversity

In fact, the two main dialects of the Kurdish language Kurmanji (Northern Kurdish) and Sorani (Central Kurdish) are spoken by more than 75% of the Kurdish speakers. Whereas, in some resources, they are considered as two different languages, based on their less intelligibility and high differences [2, 10]. Besides the two main dialects, there are Zazaki and Gurani dialects but, our survey does not consider them since there is no single NLP study that covers them, and they are very less resourced.

There are some basic differences between Sorani and Kurmanji [2, 20]. In the case of gender assign, the Kurmanji dialect is more restricted to both genders (male and female) unlike Sorani one, even there are some exceptions cases even in the Kurmanji dialect.

Sorani dialect does not follow case opposition (oblique and absolute), unlike Kurmanji, which has a clear conservative of it. In Kurmanji past tense, there are full ergative transitive verbs, while, Sorani is free of that. The creation of passive and causative is also different; in Sorani, a morphological change in the verb makes them, besides, in Kurmanji formed by helper verbs [14].

3.2 Orthography (Script)

Parso-Arab, Cyrillic, and Yekgrtu – thus scattering in orthography occurred not only because of the dialect difference. Such factors as politics, geography, and culture have the main role in it [2, 10, 12]. Previously, an attempt was done to unite all the Kurdish systems of writing for all dialects called Yekgirtú, but due to the previously mentioned reasons, this was not so successful. Apart from that, each dialect uses regional scripts. For example, some small Kurmanji dialect speakers in the former Soviet Union use the Cyrillic alphabet. The remaining Kurmanji dialects use a modified Latin alphabet.

For Sorani Dialect the modified Arabic (Parso-Arabic) alphabet is used [4, 13]. The orthography varieties make it difficult to deal with all dialects with the same tools or approach; as is shown in Table 1. The capitalization absence increases the challenges of many prepress operations related to NLP, like tokenization, segmentation, and named entities recognition in the dialects which use Parso-Arabic Script [4].

Table 1. Kurdish dialects and their scripts

dialect	Script
Zazaki	Latin
Sorani	Parso-Arabic
Kurmanji	Latin
Kurmanji	Cyrillic
Gorani	Parso-Arabic
Kirmashani	Parso-Arabic

4. Dialect Base Classification of the Studies

According to our collected data, the studies done in the Kurdish Natural language processing largely covered the Sorani Kurdish dialect or Bi-dialect of Sorani and Kurmanji. This fact made our study concentrate in this direction, besides highlighting the fact that only two pure studies covered Kurmanji dialect and one partially mentioned Zazaki in a small section. This data is clarified in Fig. 1.

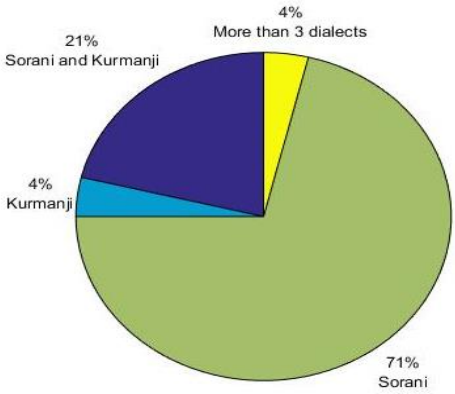

Figure 1. Dialect base distribution of studies

4.1 Kurmanji dialect-related nlp studies

Kurmanji dialect text and script could be quite familiar and less challengeable than Parso-Arabic script in so many aspects of preprocessing [5, 12]. But, according to the number of studies done on the Kurmanji dialect this fact is less affected. Many other hidden factors could be behind this scenario.

In 2010, Walther *et al.* developed a basic NLP tool for pre-processing in Kurmanji Dialect [20], by using raw corpora, non-formalized grammar reference, and lexicon data. The authors constructed a morphological lexicon and part of speech tagger for Kurmanji dialect called (Kurlex), the lexicon contains about 22,300 entries with 36 POS tagging kinds. It was 85.7% accuracy for evaluating small test data (13 sentences have 168 tokens); this accuracy could be improved by using a bigger lexicon that covers more entries as the result of other languages shown with the same method. In 2017 Gökırmak & Tayers had a useful attempt to build the dependency treebank which is the first one among Kurdish dialects [11]. The work is done on corpus over 10,000 tokens in 780 sentences collected from Wikipedia and translated stories, using 675 sentences for parser training; 75 – for testing the results; 88.3, 882, and 78.6 – for Lemma, POS, and morphological analyses using UDPipe which is the best model, by adding the result improved by 2% into the dictionary. Both studies' links and details are in Table 2.

Table 2. Kurmanji Studies

Authors	Year	Keywords
Gökırmak and Tyers [1]	2017	Kurmanji, Kurdish DT, NLP Kurdish
Walther & Sagot, & Karen[2]	2010	Kurmanji, POS, NLP, Kurdish

[1]https://www.aclweb.org/anthology/W17-6509/

[2] https://hal.archives-ouvertes.fr/halshs-00751193

4.2 Sorani dialect nlp studies

The works are done on Kurdish Sorani dialects as is shown in Fig. 1; they make up 65% of the Kurdish NLP studies. Although, the Kurdish Sorani dialect speakers are less than Kurmanji in general, and dealing with Sorani text is more complicated as already explained in Section 3 about the challenges. But we found twenty-seven studies related to Sorani NLP. Then, classifying them into some basic categories according to the aim and method of the studies, thus we can make our survey more effective. The Sorani studies' topics are:

4.2.1 Building Sorani resource

Despite the fact, that some lexicon electronic dictionaries are available and done without any documentation, our study does not present them.

Only studies that have been published as conferences or articles are presented in our survey. We classified them according to their publication's date. A brief review is as it is shown in Table 3.

Table 3. Sorani Building resource studies

Authors and work	Year	Keywords
Guatier Gerard [1]	2006	Computational linguistics, text corpus & Kurdish
Walther & Sagot[2]	2010	Sorani Kurdish. NLP
Esmaili, Elyasi, Salavati,Aliabadi & Mohammadi, Yosefi & Hakimi[3]	2013	NLP sorani, Kurdish Lemmatizer
Aliabadi, Salavati, Ahmadi& Esmaili[4]	2014	Nlp Kurdish Sorani
Esmaili, Salavati & Datta[5]	2014	Kurdish Language, Bi-Standard Languages, Test Collection, Stemming, Cross-Lingual Information Retrieval
Hosseini, Veisi & Mohammadamini[6]	2015	زبان کُردی ،پردازش زبان طبیعی6 ، ساختار تصریفی 4 کلیدواژ ههها: واژگان زایا 7 سورانی
Sardar Jaf[7]	2016	NLP, Kurdish
Salavati & Sina[8]	2018	NLP sorani, Lemmatizer
Mustafa & Tarik[9]	2018	Kurdish stemming; list of Kurdish stop words; stemming approaches
Mohammadamini& Veisi& Hosseini[10]	2019	NLP Sorani
Husseını & Mahmudi[11]	2019	زمانی کوردی، نۆرماڵکردنی دەق، ڕێنووسی کوردی، یوونیکۆد، کۆرپسی دەقی ئاسۆسافت
Husseını, Huseinı & Amini[12]	2019	زبان کُردی ،پردازش زبان طبیعی ، ساختار تصریفی کلیدواژ ههها: واژگان زایا سورانی
Husseını, Hadi & Amini & Mahmudi[13]	2019	زمانی کوردی، کوردیی ناوەندی، زمانی ستاندارد، کۆرپسی دەقی ئاسۆسافت.
Abdulrahman, Hassani & Ahmadi[14]	2019	Sorani dialect, corpus
Abdulrahman & Hassani[15]	2020	Kurdish Sorani dialect, corpus
Ahmadi, Hassani & Abedi[16]	2020	Computational Folkloristics, less-resourced languages, lyrics corpus, Kurdish
Ahmadi, Hassani[17]	2020	Morphological analysis, finite-state transducers, less-resourced languages, Kurdish
Sina Ahmadi[18]	2020	text preprocessing, stemming, tokenization, lemmatization, and transliteration

[1]https://www.institutkurde.org/en/conferences/kurdish_studies_irbil_2006/Gerard+GAUTIER.html

[2] https://halshs.archives-ouvertes.fr/halshs-00751634

[3] https://ieeexplore.ieee.org/document/6616470

[4] https://aclweb.org/anthology/papers/W/W14/W14-0101/

[5] https://dl.acm.org/citation.cfm?id=2556948&dl=ACM&coll=DL

[6] https://www.researchgate.net/publication/333856055_KSLexicon_Kurdish-Sorani_Generative_Lexicon (Persian)

[7] http://dro.dur.ac.uk/19597/

[8] https://arxiv.org/abs/1809.10763

[9] https://journals.sagepub.com/doi/10.1177/0165551516683617

[10] https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqy074/5310055

[11] https://www.academia.edu/39547697/Automated_Kurdish_Text_Normalization

[12] https://www.researchgate.net/publication/333856055_KSLexicon_Kurdish-Sorani_Generative_Lexicon

13https://www.researchgate.net/publication/333827836_Challenges_in_Standardization_of_Kurdish_Language_A_Corpus

[14] https://arxiv.org/abs/1909.11467

[15] https://arxiv.org/pdf/2004.14134.pdf

[16] https://aran.library.nuigalway.ie/handle/10379/15921

[17] https://arxiv.org/pdf/2005.10652.pdf

[18]https://www.aclweb.org/anthology/2020.nlposs-1.11/

Gautier in 2006 proposed a basic system for building Kurdish corpus. He presents a method for 100,000 words corpus [21]. His suggestion was a basic method like aConeCorde (Arabic lexicon) by Andrew Roberts for Kurdish written in Arabic letters. In 2010, Walter and Sagot developed a comparable large lexicon for Sorani Kurdish [22], collecting the raw data from different sources implementing a semi-automatic method to build the lexicon, the output SoraLex which is part of the Alexina Framework. Alexina covers some other less-resourced languages, SoraLex contains 17.600 extensions which cover 48.4% of the raw data tokens.

In 2013, Esmaili *et al.* built and collected a test set for Sorani dialect; the project is called Pewan [23]. Besides, viewing the challenges of NLP Sorani which is the less-resourced language dialect, the project aimed to build a large and net Sorani corpus and to make it available freely later for various NLP studies, like information retrieval. Pewan is part of the bigger project – the Kurdistan language processing project (KLPP2). It was built by carefully following the standard TREC's test collection construction methodology. First, they collected a large volume of documents written in Sorani, and then a powerful desktop search tool to compile a list of queries was used. Next, the authors leveraged three

widely-used open-source information retrieval systems, as well as the implementation of two well-known retrieval models to create result pools for all queries. These pools were manually assessed by authors to generate the true list of relevant documents for each query. The previous work was published in an article under the title "Towards Kurdish information retrieval" in 2014 [47]. By 2015, Hosseini *et al.* were adding their effort to the Kurdish Sorani studies under the KSLexicon title [24]. The work contained 35000 words and stems of the Sorani dialect, the summary of the study was published under the title "Kurdish Sorani Generative lexicon: وازكان زاياي زبانى كوردي سوراني" in the Persian journal using Persian language.

In 2016, Jaf suggested a semi-automatic way for the Kurdish Sorani text to unify Unicode values [4]. The approach can reduce the normalization stage problems in the preprocessing step. As it was applied to the Farsi language, there is possibility to give similar results in the related languages, like Urdu and Pashtu in the future. In 2018, Saeed *et al.* proposed an approach for Kurdish Sorani stemming [25]. The method was following normalization and stopword removal to get meaningful stems and use them for other NLP processes.There is always mismatching in this kind of tool as the Kurdish Sorani includes a huge range of stopwords and affixes. In the same year, Mustafa and Tarik proposed an approach for Kurdish Sorani stemmer to use it in information retrieval as a single stem or term decrees the dimensionality of the features [6]. Still, in 2018, Salavati and Ahmadi [26] suggested Peyv and Renus lemmatizers and word-level error correctors for Sorani Kurdish. The approach had an excellent accuracy of 96.4% when it was ran with a lexicon and 87% accuracy without a lexicon, and the lemmatization accuracy showed 86.7% accuracy. Peyv and Renus are considered the first available lemmatizer and spell-checker. The approach procedure was done based on morphological rules and n-gram Sorani language model, the data were taken from Pewan text corpus. In 2019, Husseini *et al.* had an attempt to normalization the Sorani Kurdish text [27]. The work published in the Kurdish language, concentrated on the general errors in Kurdish corpora, at the same time how to automatically deal with them. General errors, spelling, stopwords, and standard Kurdish Unicode were treated. These steps were implemented on AsoSoft corpus. The same group with

Muhamadi showed in an article the general issues related to the preprocessing steps [28]. These issues faced the Kurdish standard language too. They tried to publish a basic Sorani Kurdish version of AsoSoft which is suggested Sorani standard Kurdish as there are some mini-dialects of Sorani too. The authors tried to normalize 100000 most used words in AsoSoft by suggesting the most used forms of these words. The last version of AsoSoft corpus was published by Veisi *et al.* In this work, the authors followed standard steps to build the first Sorani text corpus with 188 million tokens [18]. Collected texts from websites, books, and magazines pipelined in a preprocessing steps to build the corpus. The authors clearly showed all the challenges and problems in each step. Also, Abdulrahman *et al.* made up a corpus depending on the Kurdish Sorani textbooks for K12 class [52]. It contains 693,800 tokens classified into 12 subjects; they listed the subjects under their sentences and token numbers.

In 2020, a new corpus development by Abdulrahman & Hassani was designed for Kurdish Sorani [53]. They used unsupervised machine learning "Punkt" as a method for segmenting the corpus. It was a part of the Kurdish BLARK project. Another direction of corpus efforts was done by Ahmadi *et al.* in 2020 [54], collecting the Kurdish folk lyrics and transcripts in Sorani. It contained 49,582 tokens for 162 songs. In 2020, Hassani continued to conduct "Towards Finite-State Morphology of Kurdish". By extracting morphological roles, he tried to analyze Kurdish Sorani words, that can be used in information retrieval or even in more advanced KNLP tools [55].

Ahmedi 2020 [56] introduced a language processing toolkit to handle the diversity of Kurdish dialects efficiently. It mainly included tokenizer and translation.

4.2.2 Sorani translation and speech-related works

In this part, we review studies about Sorani Kurdish translation and speech-related works as a part of the Kurdish NLP, ignoring the dictionaries and lexicons that are working on a word or token-level translators.

In 2009, Daneshfar *et al.* had one of the earliest studies in Kurdish Natural Language Processing (KNLP) [29]. They developed a system for Kurdish, Text-to-Speech (TTS). The system used an allophone to concatenate synthesis. The results were tested according to their

intelligibility, naturalness, and quality of speech. It was tested by twenty volunteers, they gave scores of 2.31, 2.71, and 2.39 as average out of 5 according to the test criteria. After two years, Hassani and Karim researched Kurdish Text to speech (KTTS) [30]. They developed a system depending on the concatenative synthesis method. They tested their speech system according to intelligibility, naturalness, and the comparison between them. According to the researchers, the results were acceptable as the testing was done by three different groups of listeners.

In 2017, Taher built the first Sorani Kurdish machine translator inKurdish [31]. The proposed tool inKurdish translates English to Kurdish following very simple methods of the direct meaning of part of speech (POS). After getting 50 different samples, they evaluated the efficiency of their system both computationally and linguistically using Asiya toolkit.

In the same year, Kaka-Khan proposed a Sorani Kurdish Chatbot [32]. The study used free open source (pandorabots) to build Kurdish conversation chatbot. Although the study is beginner and simple but explained some challenges with issues that face this kind of project and their future possible solutions. After one year, Kaka-Khan proposed a rule-based English to Kurdish machine translation [33]. The proposed rule-base showed noticeable improvements compared with the previous machine translator while the Sorani Kurdish language is still poorely resourced. In 2019, Qadar and Hassani had a try to develop a system for Kurdish simple speech recognition through converting simple sentences speech to text [51]. They trained the system depending on the primary school textbook. Depending on the dictionary, phone set, and transcription, it tried to understand 200 Sorani Kurdish sentence samples. In 2020, Ahmadi and Masoud [60] addressed the main issues in creating a machine translation system for the Kurdish language, with a focus on the Sorani Kurdish-English translation. In the same year, Kamal and Hassani proposed a project which aimed to develop the necessary data and tools to process the sign language for Sorani as one of the spoken Kurdish dialects [61]. The result was the development of a dataset in HamNoSys and it was a corresponding SiGML form for the Kurdish Sign lexicon. Table 4 shows the works based on translation and speech studies.

Table 4. Translation in Sorani

Authors and work	Year	Keywords

Daneshfar & Barkhoda[1]	2009	Text to Speech System; Kurdish language; Concatenative TTS; Allophone and Kurdish TTS
Hassani & Karim [2]	2011	Speech Kurdish Sorani NLP
Kaka-Khan [3]	2017	AI, AI Markup Language, Chatbot, Pandorabots
Kaka-Khan & Taher [4]	2017	NLP, Machine Translation (MT), Kurdish, Asiya Toolkit, inkurdish translator, BLEU, NIST, METEOR
Kaka-Khan [5]	2018	Apertuim, Inkurdish, Machine Translation, Morphological, Rule-based Machine Translation
Qader & Hassani [6]	2019	Speech Kurdish, STT, Sorani
Ahmadi & Masoud [7]	2020	Sorani Kurdish-English translation
Kamal & Hassani[8]	2020	Sign Language, Kurdish Language Processing, Kurdish to Sign, HamNoSys, SiGML

[1]https://ieeexplore.ieee.org/document/6297277

[2] https://www.aclweb.org/anthology/W16-4812

[3] http://juhd.uhd.edu.iq/journals/images/Vol1_No4/393-397.pdf

[4] https://scialert.net/abstract/?doi=itj.2017.27.34

[5] https://link.springer.com/chapter/10.1007%2F978-3-319-59463-7_19

[6]https://www.researchgate.net/publication/321421117_Plagiarism_Detection_System

[7] https://link.springer.com/article/10.1007/s42044-018-0007-4#citeas

[8] http://eprints.ukh.edu.krd/81

4.2.3 Sorani semantic studies

Fetching information from a text or speech is considered the advanced level of natural language processing for any natural language. Kudish case is in the first decade, according to collected studies. At the beginning of 2012, Mohammed *et al.* published a work, which was the first study related to the classification of Sorani texts into four categories: art, economy, politics, and sport [34]. The procedure of the study was simply by normalizing and tokenizing the texts, then testing the documents according to N-gram categorized texts.

In 2016, there were two studies that can be classified as follows:

The first one was done by Shervin Malmasi; he tried to automatically identify the subdialects in Sorani Kurdish [2]. The support vector machine was used for the task. more than 200000 sentences from different sources were collected, depending on the different n-gram features of the texts. The classifier worked with 96% accuracy, distinguishing subdialects' differences of news sources in Iran and Iraq. It was the first experimental method for this task.

The second study was implemented by Abdulla and Hama; they based on Naïve Bayes classifier to identify negative or positive ideas in

Social Networks for Kurdish Sorani [35]. The best accuracy they got was 66% with a 0.72 F-score.

In 2017, Rashid *et al.* applied a robust classification system for Sorani Kurdish files [36]. The tool tried to classify any document given to it into one of eight classes: sports, religions, economics, arts, socials, styles, education, and health. They used support vector machine and decision tree classifiers. The preprocessing step of stemmer increased the efficiency of the system, getting 83.2% and 93.1% for the decision tree and SVM classifier respectively. The same group of authors one year later used KDC-4007 dataset for classification. The data set is compatible with SVM, NB, and DT classifiers. The best result was 91.03 by SVM and the worst one was for the NB classifier – 5.1% [16]. The same year Wakil *et al.* proposed a system for Sorani Kurdish plagiarism detection [48]. The system depended on the n-gram method for the detection of words, phrases, and paragraphs. The system connected to the local sources and google search to implement its work.

In 2018, Saeed *et al.* tried another approach for Sorani Kurdish classification using Reber Stemmer [25]. The work was simply classification of Kurdish documents into some categories by taking the minimum possible stem and without stemming then the compared results. In this procedure, again SVM and DT were used. The same group of researchers conducted another improvement of classification using porter stemmer with tree data structure [37]. The performance of the classification was taken in different scenarios like considering stopwords besides some other stemmer details with explaining the efficiency of their classifications. Table 5 shows above studies.

Table 5. Sorani sentiment analysis

Authors and work	Year	Keywords
Mohammed, Zakaria, Omar & Albared[1]	2012	N-Gram, text categorization, Indo-European languages family, Kurdish Sorani, Unicode, Dice Measure of similarity, text representation.
Malmasi [2]	2016	Nlp kurdish Sorani
Shaltooki, Mzhda[3]		Sentiment Analysis; Kurdish Sentiment; Naive Bayes Classifier
Tarik, Arazo & Ari [4]	2017	Documents classification, Kurdish stemming, machine learning algorithm, information retrieval
Tarik , Arazo & Ari [5]	2017	SVM Text Classification Term FT Document DT Classifier
Wakil, Ghafoor, Abdulrahman	2017	Plagiarism Detection, Plagiarism Detection System,

& Tariq[6]		N-Gram, Kurdish Language, Theft.
Ari, Tarik,Mustafa, Al-Rashid & K. Al-Salihi [7]	2018	Kurdish text classification Stemming SVM DT
Ari, Tarik, Arazo, Polla & Birzo [8]	2018	Kurdish text classification, Porter's stemmer algorithm, Stemming, Tree data structure

[1]https://ieeexplore.ieee.org/document/6297277
[2] https://www.aclweb.org/anthology/W16-4812
[3] http://juhd.uhd.edu.iq/journals/images/Vol1_No4/393-397.pdf
[4] https://scialert.net/abstract/?doi=itj.2017.27.34
[5] https://link.springer.com/chapter/10.1007%2F978-3-319-59463-7_19
[6]https://www.researchgate.net/publication/321421117_Plagiarism_Detection_System
[7] https://link.springer.com/article/10.1007/s42044-018-0007-4#citeas
[8] http://eprints.ukh.edu.krd/81

4.3 Dialectal (sorani-kurmanji) studies

Some studies in Kurdish natural language processing tried to cover as wide as possible range of the language and dialects as the standard Kurdish language is not available yet. The study of the Kurmanji and Sorani Kurdish dialects is one of the possible choices while both of the dialects together cover more than 75% of the Kurdish language speakers.

4.3.1 Building resources

The beginning steps in building Kurdish resources that included Kurmanji and Sorani dialects together were made by Gautier in 1996 – the Dirêjî Kurdî (lexicographic environment for Kurdish language using 4th Dimension) [39]. The work is considered as a lexicon and software environment for developing a future translator or any other Kurdish NLP at that time. Two years after his first study about Kurdish NLP, Gautier published another work to build the first Kurdish Corpus [40]. The author clarified the basic issues which face building a Kurdish corpus and the basic procedure required, besides the importance of having a good corpus.

In 2012, Esmaili analyzed the most important challenges faced by Kurdish natural language processing in both Kurmanji and Sorani [5]. The study purified the dialect differences as it's more complicated than transliteration but less than translation. Therefore to have any strong NLP study these differences should be taken into consideration.

In 2013, Esmaili and Salavati statistically focused on some morphological, phonological, and orthographic differences in Kurmanji and Sorani dialects [14]. The proposal of the Kurdish corpus Pewan was another income of their study. By 2014, Aliabadi suggested a Kurdish prototype KurdNet. He highlighted the main challenges faced by the

KurdNet, based on a stable plan explained using it in the Natural Language Processing system developed [5, 40].

In 2017, Hassani proposed a procedure for Noun extraction from Kurdish text [42]. The approach of the name entity recognition basically was about personal names identifier. The tool gave more than 95% precision. The same year, Hassani suggested BLARK (Basic Language Resource Kit), as a solution for Kurdish language diversity in dialects and challenges to be used in various natural language processing and computational linguistic processes [47]. The author expected to compare the use of BLARK in other languages as while Kurdish makes a huge step. In 2019, Ahmadi *et al.* proposed tri-dialect Kurdish lexicography [50], which covered Sorani, Kurmanji, and Hawrami. The work reviewed the lexicography state of three Kurdish dialects, as they have 60% of the Kurdish lexicon resources. The work depended on three grammar books dialect related, with 4172, 5683, and 1184 words in each dialect respectively.

In the recent 2020 works by Ahmadi [57], an approach was proposed for both Kurmanji and Sorani tokenizing depending on morphological and lexicon analysis. Ahmadi *et al.* 2020 proposed a corpus containing 12,327 translation pairs in Sorani and Kurmanji dialects. Also, he provided 1,797 and 650 translation pairs in English-Kurmanji and English-Sorani [59]. These studies are listed in Table 6 below.

Table 6. Multi-dialect or more building resources

Authors and work	Year	Keywords
Gautier [1]	2006	Kurdish, multiscript, database
Gautier [2]	2008	Kurdish, multiscript, corpus, lexicography
Esmaili [3]	2012	NLP Sorani, Kurmanji
Esmaili & Salavati [4]	2013	Not mentioned
Aliabadi [5]	2014	Kurdish NLP
Hassani [6]	2017	Proper Noun Recognition, Named Entity Recognition, Information Extraction, Natural Language Processing, Kurdish
Hassani [7]	2017	Kurdish BLARK Language tools Computational linguistics Natural language processing
Ahmadi,Hassani & McCrae[8]	2019	Kurdish;e-lexicography; less-resourced languages; machine-readable dictionary
Ahmadi [9]	2020	Tokenize, Sorani-Kurmanji, preprocessing.
Ahmadi et al[10]	2020	Kurmanji Sorani Kurdish corpus.

[1]file:///D:/NLP-All/Kurdish/nlp-new/Direji_Kurdi_a_lexicographic_environment_for_Kurdi.pd

[2] http://ggautierk.free.fr/e/icem_98.htm

[3]https://arxiv.org/abs/1212.0074

[4] https://ieeexplore.ieee.org/document/6616470

[5] https://www.semanticscholar.org/paper/Semi-Automatic-Development-of-KurdNet

[6]https://www.researchgate.net/publication/316511715_A_Method_for

[7] https://link.springer.com/article/10.1007/s10579-017-9400-0

[8] https://www.library.ucg.ie/handle/10379/15513

[9] https://www.aclweb.org/anthology/2020.vardial-1.11/

[10] https://arxiv.org/pdf/2010.01554.pdf

4.3.2 Multidialectal processing studies

Studies that cover more than one Kurdish dialect in our classification are six studies as shown in Table 7. Almost all of them covered Kurmanji and Sorani dialects. Furthermore, in this part beyond preprocessing approaches are highlighted text to speech, dialect identification, and transliteration.

In 2009, Barkhoda *et al.* developed and tested a Text To Speech system for Kurdish [43]. The authors tested syllable, allophone, and diphone synthesis in their system. The results explained the diphone-based system were more efficient. In 2016, Hassani and Medjedovic in an Automatic Kurdish dialect identification study applied a support vector machine (SVM) for Kurdish dialect identification and classification [12]. A list of proper words in each dialect was used for training the system. Identifying the dialect of Kurmanji and Sorani by the system was 92% and 91% respectively.

Hassani in 2017 developed an interdialect machine translation from Sorani to Kurmanji and vice versa. It had 63% and 71% understandability respectively [44]. The used method was basically word-to-word translation. The same year, Hassani and Hamid added an Artificial Neural network (ANN) to the Kurdish dialect identification [45]. For a short text or a single sentence, the ANN approach got 99% and 96% accuracy for Kurmanji and Sorani respectively, while for the long piece of text no significant improvement was shown compared with the previous methods. Ahmedi in 2019 suggested another method for transliteration of Kurdish dialects under the Wergor transliteration system name [46]. The efficiency based on rule transliteration reached 99%, in some cases availability of little mistakes with bigger data sets were easily overcome.

In 2020, Ahmadi [58] tried to build a corpus mainly for Zaza and Gorani dialects. The work contained over 1.6M and 194k word tokens, respectively. Table 7 illustrates these studies.

Table 7. Multi-dialect processing studies

Authors and work	Year	Keywords
Barkhoda, Azami, Bahrampour & Shahryari [1]	2009	speech synthesis; text analysis; Kurdish language; allophone TTS systems; syllable TTS systems; diphone TTS systems
Hassani & Dzejla [2]	2010	Dialect identification, NLP, Kurdish language, Kurmanji, Sorani
Hassani [3]	2017	Sorani, Kurmanji. NLP
Hassani & Hamid [4]	2017	Dialect Classification, NLP, ANN, Machine Learning, Kurdish Dialects.
Ahmadi [5]	2019	Transliteration, rule-based approach, Kurdish, less-resourced language processing
Ahmadi[6]	2020	Corpus Kurdish, Zaza, Gorani.

[1]https://ieeexplore.ieee.org/document/5407540
[2] https://www.researchgate.net/publication/295093908_Automatic_Kurdish_Dialects_Identification
[3] https://www.researchgate.net/publication/314736645_Kurdish_Interdialect_Machine_Translation/related
[4] https://www.scitepress.org/PublicationsDetail.aspx?ID=ttUX6lbOy9g=&t=1
[5] https://dl.acm.org/citation.cfm?id=3278623

5 Synthesis and Discussion

There are presented and classified 52 research papers; all are total Kurdish NLP studies that have been analyzed and listed since 2021. The review is the first attempt related to the KNLP. Although the covered period is 1996 till 2020, more than 50% of the papers were published in the last three years (29 papers out of 52).

Till 2009 from our starting point there were only three works related, depending on that 2009 considered the real starting point for KNLP as it is illustrated in Fig. 2. Dialect-based classification indicates that 66.66% of the papers are purely related to the Sorani dialect works which is a dialect of less than 30% of the Kurdish population. In contrast, 4.44% of works represent the dialect of 65% of the Kurdish speakers, Kurmanji. However, both dialects' studies together are 24.44%. The remaining 4.44% of studies covering more than three dialects included Kurmanji and Sorani.

Generally from the aforementioned statistics, the majority, i.e., 95% of the studies, covered the main two dialects together (47 works out of 52). The mentioned above is clearly illustrated in Fig. 1 and 2.

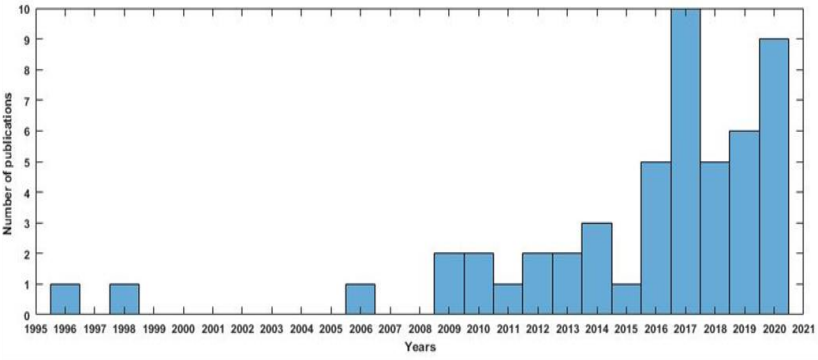

Figure 2. Studies according to publication year

In Fig. 3 the analysis of the nationality of the authors, as the country-based studies, is presented. Three trends are obvious: 1) the works done represented by authors from Iraq and Iran make 80% of the studies (36 out of 52 works). This may give us a clear clue to the gap between Kurmanji and Sorani studies as mainly the Sorani speakers are there. 2) The Kurmanji studies alone make 4.44%, where the main ratio is of the authors' publications from Turkey. This can be the remarkable reason for poor Kurmanji KNLP works, although they represent more than 65% of Kurdish speakers. 3) The remaining 15% of the studies scattered among France, the US, the UK, with a remarkable 24% from France. This can be illustrated from the support that Kurdish languages studies are got by different institutes since old times.

In the aspect of the building resource from Table 6 and 7, the purely related works are 33.33% (with 15 works out of 52). This ratio is divided as 28.88% and 4.44% respectively between Sorani and Kurmanji dialects. Despite this fact, most of the other works have prepared their own data through preprocessed steps from scratch as the Kurdish language is one of the less-resourced languages. In addition, the morphological complexity for the Kurdish language, especially for the Sorani dialect which uses the updated Fars-Arabic alphabet, makes stemming, lemmatization, and POS processes obligatory steps to handle Sorani dialect NLPs.

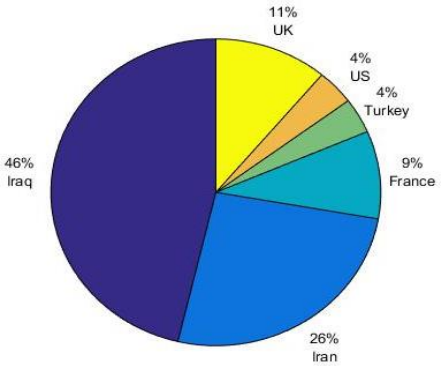

Figure 3. Studies according to authors' belonging countries

For the trend related to the number of authors, the noticeable point is that 32% (17out of 52) of the works are personal effort with a single author, which is the highest ratio of the works compared to none single authors per works. Fig. 4 explains this point. This issue faced by KNLP studies can consider one of the reasons behind quality and quantity of the works, in the other expression, the lack of group works and organizational efforts among the institutes.

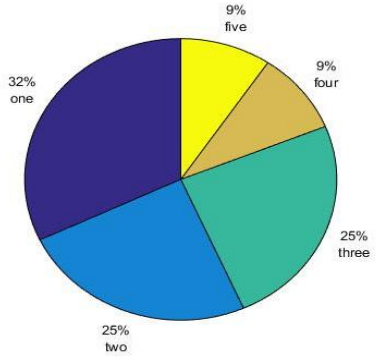

Figure 4. Number of authors per study

Finally, the observation about the quality and the technical tools deployed in the works is quite apparent. For the quality part, the topic base compression is the most used in similar studies [62, 63]. In our work on KNLP, the scatter of the studies topic and direction base make the efficiency compression according to their result fairly impossible now. The preprocessing works almost depended on semi-automatic systems as most of the studies deal with Sorani dialect and the use of Parso-Arabic script is one of its properties. This made dealing with tokenizing, lemmatization, stemming, and POS extremely difficult by fully automatic systems. Some open sources tools and developed approaches were used in the translation and transliterations works. However, the studies that included sentiment analysis and classification used the traditional approaches of NV, SVM, DT, and BDT, but in a few cases, the ANN was deployed [42].

6 Conclusion and Perspectives

All Kurdish NLP studies are analyzed and classified in our work. This includes 52 papers and works covered in available studied areas of KNLP. The review originality was conducted by associating each work to its resource, tool, and publicity available. From synthesis and analysis of the present works, we deduce the fact that only one work has approached the dialect of Zazaki, while Kirmashani and Hawramani dialects were not covered in the KNLP.

In most cases, many KNLP issues have not been discussed, besides, many others just opened. Most of the works were on Kurdish text whereas the Kurdish speech studies were very few in comparison. Although most of the KNLP studies were applied to the Sorani and Kurmanji dialects, even these dialects have not got adequate covering. The language is still less-resourced compared with English or even Arabic languages.

Our survey shared many research hints and questions: for the dialect part, is it better to cover more than one Kurdish dialect in the works or take each dialect apart, while there are the script and morphological differences among the dialects? For the technical part, should the building resources be pre-step in each work, or depending on the built resources suffusion? For this purpose, the manual method or automatic techniques perform better. The associated works between authors and institutes are

few; this affects the quality and the quantity of the studies related to the KNLP. While techniques and approaches like deep learning are deployed in the NLP, why do almost all the KNLP studies use traditional methods like SVM, NB, etc., although, more than 50% of the studies were conducted in the last three years?

The idea behind this survey is to collect and classify all the KNLP studies, introduce the research community to the level of the studies and the methods used. By analyzing the works and their directions, some reason behind the poor and barely covered areas is shown. In the future, for performing a new survey, the tools, methods, and resources used can be detailed, demonstrating the power and weakness of each step. Also, explaining the parallel tools used in other languages helps the researchers to have a direct and easy view of KNLP.

References

[1] Steedman, Mark. "*Natural language processing*." Artificial intelligence. Academic Press, 1996. 229-266.

[2] Malmasi, S. *Subdialectal differences in sorani kurdish*. In Proceedings of the third workshop on nlp for similar languages, varieties and dialects (vol: vardial3) (pp. 89-96) (2016, December).

[3] Republic of Iraq. "*Constitution of the Republic of Iraq*", available at: https://www.refworld.org/docid/454f50804.html, 15 October 2005.

[4] Jaf, S. *A simple approach to unify ambiguously encoded Kurdish characters.* In Proceedings of the International Conference Computational Linguistics in Bulgaria (2016, September). (pp. 86-94). Institute for Bulgarian Language, Bulgarian Academy of Sciences.

[5] Esmaili, K. S. *Challenges in Kurdish text processing*. arXiv preprint (2012). vol:1212.0074

[6] Aliabadi, P., Ahmadi, M. S., Salavati, S., & Esmaili, K. S. *Towards building kurdnet, the Kurdish wordnet.* In Proceedings of the Seventh Global Wordnet Conference (2014, January). (pp. 1-6).

[7] Mohammed, B. O. *Handwritten Kurdish character recognition using geometric discertization feature.* International Journal of Computer Science and Communication, (2013). vol:4,pp. 51-55.

[8] Windfuhr, G. (Ed.). *The Iranian Languages*. (2009). Psychology Press.

[9] Austria: Federal Ministry of the Interior, *The Kurds: History –Religion – Language - Politics, November.* 2015, available at: https://www.refworld.org/docid/568cf9924

[10] Haig, G., & Matras, Y. *Kurdish linguistics: a brief overview. STUF-Language Typology and Universals*, (2002). *Vol.* 55(1), pp. 3-14.

[11] Gökırmak, M., & Tyers, F. *A dependency treebank for Kurmanji Kurdish*. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017) (pp. 64-72).

[12] Hassani, H., & Medjedovic, D. *Automatic Kurdish dialects identification.* Computer Science & Information Technology, (2016). Vol. 6(2), 61-78.

[13] Littell, P., Mortensen, D. R., Goyal, K., Dyer, C., & Levin, L. *Bridge-language capitalization inference in western iranian: Sorani, kurmanji, zazaki, and tajik.* In Proceedings of the Tenth International Conference on Language Resources and Evaluation (vol. LREC'16) (2016, May). (pp. 3318-3324)

[14] Esmaili, K. S., & Salavati, S. *Sorani kurdish versus kurmanji kurdish: An empirical comparison.* In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers) (2013, August). (pp. 300-305).

[15] Zahedi, K., & Mehrazmay, R. Definiteness in sorani kurdish and english. Dialectologia: revista electrònica, (vol.7) (2011)., 129-157.

[16] Rashid, T. A., Mustafa, A. M., & Saeed, A. M. *Automatic Kurdish text classification using KDC 4007 dataset.* In International Conference on Emerging Internetworking, Data & Web Technologies (2017, June). (pp. 187-198). Springer, Cham.

[17] Nerwiy, H. K. T. The Republic of Kurdistan, 1946 (Doctoral dissertation, Faculty of the Humanities, Leiden University) (2012).

[18] Veisi, H., MohammadAmini, M., & Hosseini, H. *Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus.* Digital Scholarship in the Humanities, (2020). vol.35(1), 176-193.

[19] Schahbasi, A., Vogl, M., Webinger, P., Schrott, T., & Bauer, S. *The Kurds: history-religion- language-politics.* (2015).

[20] Walther, G., Sagot, B., & Fort, K. *Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish.* (2010, September).

[21] GAUTIER, *Computerised text corpus in Kurdish*, World Congress of KURDISH STUDIES. 2006.

[22] Walther, G., & Sagot, B. *Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish.* (2010).vol. 128 pp. 1-26.

[23] Esmaili, K. S., Eliassi, D., Salavati, S., Aliabadi, P., Mohammadi, A., Yosefi, S., & Hakimi, S. *Building a test collection for Sorani Kurdish.* ACS International Conference on Computer Systems and Applications (AICCSA) (2013, May). (pp. 1-7). IEEE.

[24] Hosseini, Hawre & Veisi, Hadi & Mohammadamini, Mohammad. *KSLexicon: Kurdish-Sorani Generative Lexicon*. (2015).

[25] Saeed, A. M., Rashid, T. A., Mustafa, A. M., Agha, R. A. A. R., Shamsaldin, A. S., & Al-Salihi, N. K. *An evaluation of Reber stemmer with longest match stemmer technique in Kurdish Sorani text classification.* Iran Journal of Computer Science, (2018). Vol. 1(2), pp.99-107

[26] Salavati, S., & Ahmadi, S. *Building a Lemmatizer and a Spell-checker for Sorani Kurdish*. (2018). Vol. arXiv preprint arXiv:1809.10763

[27] Mahmudi, Aso & Veisi, Hadi & Mohammadamini, Mohammad & Hosseini, Hawre. Automated Kurdish Text Normalization خاوێن کردنی ئۆتۆماتیکی دەقی کوردی. (2019).

[28] Mohammadamini, Mohammad & Veisi, Hadi & Mahmudi, Aso & Hosseini, Hawre. *Challenges in Standardization of Kurdish Language: A Corpus-based approach.* (2019).

[29] Daneshfar, F., Barkhoda, W., & Azami, B. Z. *Implementation of a Text-to-Speech System for Kurdish Language.* In Fourth International Conference on Digital Telecommunications. (2009, July). (pp. 117-120). IEEE.

[30] Hassani, H., & Kareem, R. *Kurdish text to speech (KTTS).* In Tenth International Workshop on Internationalisation of Products and Systems (2011). (pp. 79-89).

[31] Taher, F. J. *Evaluation of inkurdish Machine Translation System.* Journal of University of Human Development, . (2017). 3(2), 862-868.

[32] Kaka-Khan, K. M. *Building Kurdish Chatbot Using Free Open Source Platforms.* UHD Journal of Science and Technology, (2017). Vol. 1(2), 46-50.

[33] Kaka-Khan, K. M. English to Kurdish Rule-based Machine Translation System. (2018).

[34] Mohammed, F. S., Zakaria, L., Omar, N., & Albared, M. Y. Automatic *Kurdish Sorani text categorization using N-gram based model.* In 2012

International Conference on Computer & Information Science (ICCIS) (2012, June). (Vol. 1, pp. 392-395). IEEE.

[35] Abdulla, S., & Hama, M. H. *Sentiment Analyses for Kurdish Social Network Texts using Naive Bayes Classifier.* Journal of University of Human Development, (2015). vol.1(4), 393-397.

[36] Rashid, T. A., Mustafa, A. M., & Saeed, A. *A robust categorization system for Kurdish Sorani text documents. Information Technology Journal*, (2017). Vol.16(1), 27-34.

[37] Mustafa, A. M., & Rashid, T. A. *Kurdish stemmer pre-processing steps for improving information retrieval.* Journal of Information Science, (2018). Vol.44(1), 15-27.

[38] Saeed, A. M., Rashid, T. A., Mustafa, A. M., Fattah, P., & Ismael, B. *Improving Kurdish Web Mining through Tree Data Structure and Porter's Stemmer Algorithms.* UKH Journal of Science and Engineering, (2018). Vol. 2(1), 48-54.

[39] Gautier, G. *Dirêjî Kurdî: a lexicographic environment for Kurdish language using 4th Dimension*® (1996, April).

[40] Gautier, G. *Building Kurdish large corpus an overview of technical problems*, In proceedings ICEMCO 1998.

[41] Aliabadi, P. *Semi-Automatic Development of KurdNet, The Kurdish WordNet.* In Proceedings of the ACL 2014 Student Research Workshop (2014, June). (pp. 94-99)

[42] Hassani, H. *A Method for Proper Noun Extraction in Kurdish.* In 6th Symposium on Languages, Applications and Technologies (2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[43] Barkhoda, W., ZahirAzami, B., Bahrampour, A., & Shahryari, O. K. *A comparison between allophone, syllable, and diphone based TTS systems for Kurdish language.* In 2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (2009, December). (pp. 557-562). IEEE.

[44] Hassani, H. *Kurdish interdialect machine translation.* In Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial) (2017, April). (pp. 63-72).

[45] Hassani, H., & Hamid, O. H. *Using Artificial Neural Networks in Dialect Identification in Less-resourced Languages.* (2017).

[46] Ahmadi, S. *A rule-based Kurdish text transliteration system.* ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), (2019). vol. 18(2), pp.1-8.

[47] Hassani, H. *BLARK for multi-dialect languages: towards the Kurdish BLARK. Language Resources and Evaluation*, (2018). Vol.52(2),pp.625-644.

[48] Esmaili, K. S., Salavati, S., & Datta, A. *Towards Kurdish information retrieval.* ACM Transactions on Asian Language Information Processing (TALIP), (2014). Vol.13(2), pp.1-18

[49] Wakil, K., Ghafoor, M., Abdulrahman, M., & Tariq, S. *Plagiarism Detection System for the Kurdish.* International Journal of Information Technology and Computer Science. (2017). Vol.12. pp. 64-71. 10.5815/ijitcs.2017.12.08.

[50] Ahmadi, S., Hassani, H., & McCrae, J. P. *Towards electronic lexicography for the Kurdish language.* In Proceedings of the sixth biennial conference on electronic lexicography (eLex). (2019, October). vol. eLex 2019.

[51] Qader, A., & Hassani, H. *Kurdish (Sorani) Speech to Text: Presenting an Experimental Dataset.* (2019). Vol. arXiv preprint arXiv:1911.13087.

[52] Abdulrahman, R. O., Hassani, H., & Ahmadi, S. *Developing a Fine-Grained Corpus for a Less-resourced Language: the case of Kurdish.* (2019). Vol. arXiv preprint arXiv:1909.11467

[53] Omer Abdulrahman, R., & Hassani, H. *Using Punkt for Sentence Segmentation in non-Latin Scripts: Experiments on Kurdish (Sorani) Texts*. (2020). Vol. arXiv, arXiv-2004.

[54] Ahmadi, S., Hassani, H., & Abedi, K. *A corpus of the Sorani Kurdish folkloric lyrics.* In Proceedings of the 1st Joint Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL) Workshop at the 12th International Conference on Language Resources and Evaluation (LREC). National University of Ireland Galway. (2020, May).

[55] Ahmadi, S., & Hassani, H. *Towards Finite-State Morphology of Kurdish.* (2020). Vol. arXiv preprint arXiv:2005.10652.

[56] Ahmadi, S. *KLPT–Kurdish Language Processing Toolkit.* In Proceedings of Second Workshop for NLP Open Source Software (2020, November). (NLP-OSS) (pp. 72-84).

[57] Ahmadi, S. *A Tokenization System for the Kurdish Language.* In Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects(2020, December). (pp. 114-127).

[58] Ahmadi, S. *Building a Corpus for the Zaza–Gorani Language Family*. In Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects (2020, December). (pp. 70-78).

[59] Ahmadi, S., Hassani, H., & Jaff, D. Q. *Leveraging Multilingual News Websites for Building a Kurdish Parallel Corpus*. (2020). Vol. arXiv preprint arXiv:2010.01554.

[60] Ahmadi, S., & Masoud, M. Towards Machine Translation for the Kurdish Language. (2020). Vol.arXiv preprint arXiv:2010.06041.

[61] Kamal, Z., & Hassani, H. *Towards Kurdish Text to Sign Translation.* In Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives (2020, May). (pp. 117-122).

[62] Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. *Arabic natural language processing: An overview.* Journal of King Saud University-Computer and Information Sciences, (2021). Vol.33(5), pp.497-507.

[63] Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. *A comprehensive survey of arabic sentiment analysis*. Information processing & management, (2019). Vol. 56(2), pp. 320-342.

Ashti Afrasyaw JAF[1], Sema Koç Kayhan[2]

[1] University of Gaziantep, Turkey, Department of Electrical and Electronics.
University of Garmian, Iraq, Department of Computer Sciences
E-mail: ashti.a@garmian.edu.krd

[2] University of Gaziantep, Turkey, Department of Electrical and Electronics.
E-mail: skoc@gantep.edu.tr

Contents

Firma poligrafică „VALINEX" SRL,
Chişinău, str. Florilor, 30/1A, 26B
tel./fax 43-03-91
e-mail: info@valinex.md
http: \\www.valinex.md

Bun de tipar 25.10.20 21
Coli editoriale 11,82. Coli de tipar conv. 12,44.
Format 60x84 1/16. Garnitură „Times".
Hirtie ofset. Tirajul 150