INNOVATIVE INFOTECHNOLOGIES FOR SCIENCE, BUSINESS AND EDUCATION

Vol. 1(12) 2012

CONTENTS

IITSBE, **Vol. 1(12) 2012**

Section: Advanced Programming

3-6 COMPARISON OF SHIFT SEQUENCE BASED AND SIMULATED ANNEALING METHODS FOR HIGHLY CONSTRAINED MEDICAL STAFF ROSTERING PROBLEMS

$$f = \exp\left[\frac{Current_{Cost} - New_{Cost}}{T}\right]$$

Mindaugas Liogys

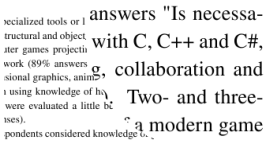
Section: IT - Business Solutions

7-13 INFORMATION TECHNOLOGIES IN INVESTMENT OPERATIONS

Jonas Žaptorius, Narghiza Sulaymonova

Section: Education

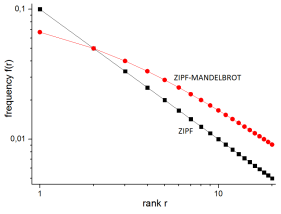
14-16 NEW TRENDS IN INFORMATICS STUDY PROGRAMMES



Eugenijus Valavičius, Violeta Jadzgevičienė

Section: Review

17-26 ZIPF AND RELATED SCALING LAWS. 2. LITERATURE OVERVIEW OF APPLICATIONS IN LINGUISTICS



Giedrė Būdienė, Alytis Gruodis

Comparison of Shift Sequence Based and Simulated Annealing Methods for Highly Constrained Medical Staff Rostering Problems

Mindaugas Liogys [a]
Vilnius University, Institute of Mathematics and Informatics, Akademijos st. 4, Vilnius, Lithuania

Received 1 June 2012, accepted 7 July 2012

Abstract. The aim of this article is to compare two different approaches (simulated annealing and shift sequence based method) used in solving medical staff rostering problem. During comparison stage two dimensions were considered: roster quality and roster building time. Test results showed that simulated annealing method is more efficient than shift sequence based method in both ways - builds a better roster in shorter time.

Citations: Mindaugas Liogys. Comparison of Shift Sequence Based and Simulated Annealing Methods for Highly Constrained Medical Staff Rostering Problems – *Innovative Infotechnologies for Science, Business and Education,* ISSN 2029-1035 – **1(12)** 2012 – Pp. 3-6.

Keywords: Shift Sequence Based Methods; Simulated Annealing Methods; Medical Staff Rostering.
Short title: Comparison of methods.

Introduction

The common objective of medical staff rostering problem is to produce rosters with a balanced workload as well as to satisfy individual preferences as much as possible.

Many researchers are interested in medical staff rostering problem. Large number of articles have appeared presenting different approaches to this problem [1-9]. Still, not many comparisons have been carried out and they focus on the comparison of two or more approaches developed by the same authors [10]. This article is devoted to compare two different approaches (shift sequence based method and simulated annealing) in solving medical staff rostering problems.

1. Problem Formulation

The problem is that of creating monthly schedules for cardiologists at a major Lithuania hospital. These schedules have to satisfy working contracts and meet as far as possible cardiologists' requests. Working contract regulations usually are called hard constraints and personal preferences - soft constraints [9]. Hard constraints and software constraints that are considered in this problem are listed in the Table 1 and Table 2 respectively.

Table 1. Hard Constraints

1.	The shift coverage requirements must be fulfilled.
2.	After night shift must be at least for 24 hours rest time.
3.	Duty shift must be assigned only on weekends.
4.	Cardiologist cannot be assigned to different assignments at the same time.
5.	Only duty shifts can be assigned on weekends.

Hard constraint N1 states that the total number of shifts on certain days must satisfy the coverage requirements. Hard constraint N2 states that there must be at least 24 hours time difference between night shift and any other shift. Hard constraint N3 states that duty shifts must be assigned only on weekends or on bank holidays. Hard constraint N4 states that if the cardiologist has more than one skill, his or her assignments must not overlap. Hard constraint N5 states that no morning, day and night shifts are allowed to be assigned on weekends. If any of these hard constraints is not satisfied then created roster is considered as improper.

Soft constraints (see Table 2) must not necessarily be satisfied; however, violations of soft constraint are penalized. Sum of penalties defines quality of roster: if the lesser sum is obtained it means the roster of higher quality is prepared. Objective of solving such problems is to minimize objective function [3]:

Table 2. Soft Constraints

1.	Maximum number of shift assignments.
2.	Maximum number of consecutive work days.
3.	Minimum number of consecutive work days.
4.	Maximum number of consecutive non-working days.
5.	Minimum number of consecutive non-working days.
6.	Maximum number of a certain shift worked.
7.	Maximum number of consecutive working weekends.
8.	Maximum number of working weekend in a month.
9.	Requested days off.
10.	Requested days on.
11.	Requested shifts on.
12.	Requested shifts off.
13.	Requested shifts for each weekday.

[a]Corresponding author, email: *m.liogys@eif.viko.lt*, cell: +370(698)87485

Table 3. Shift types

Shift label	Shift type	Time period
R1	Morning	07:30 - 15:12
R2	Morning	07:30 - 09:18
R3	Morning	09:18 - 15:12
R4	Morning	07:30 - 11:06
R5	Morning	09:18 - 14:24
R6	Morning	09:18 - 14:36
R7	Morning	11:06 - 15:12
D1	Day	15:12 - 17:00
D2	Day	15:12 - 18:48
D3	Day	17:00 - 20:36
D4	Day	15:12 - 20:36
N1	Night	18:48 - 24:00
N2	Night	00:00 - 09:12
Dt1	Duty	08:00 - 24:00
Dt2	Duty	00:00 - 08:00

$$y = \sum_{i=1}^{n} \sum_{j \in F(i)}^{m} p_{ij} x_{ij} \quad (1)$$

where n represents number of cardiologists; m - number of shift sequences; p_{ij} - cost of cardiologist i working shift sequence j; $F(i)$ - set of feasible shift sequences for cardiologist i; x_{ij} - decision variable which is equal to 1 if cardiologist i works shift sequence j, 0 - otherwise.

There are 17 different shifts available, according to various lengths of working hours for cardiologists - see Table 3. Rostering period is one calendar month.

Part of cardiologists has full time workload; part of cardiologists has part time workload; part of cardiologists has more than full time workload. There are cardiologists who have more than one skill and in order to construct correct roster have to be considered that his / her assignments do not overlap. Best case scenario is then one assignment ends and starts another for those who have several skills, i.e. no time interval between assignments on the same day.

2. Overview of methods

A solution of rostering problem consists of a collection of personal schedules for each of the cardiologists. A schedule for a cardiologist consists of shifts that usually are different in lengths and types (morning shifts, day shifts, etc.).

Table 4. Hard constraints categorized to schedule and roster constraints

N	Hard constraint	Category
1.	The shift coverage requirements must be fulfilled.	Roster
2.	After night shift must be at least 24 hours of rest time.	Schedule
3.	Duty shift must be assigned on weekends.	Schedule
4.	Cardiologist cannot be assigned to different assignments on the same time.	Schedule
5.	Only duty shifts can be assigned on weekends.	Schedule

Table 5. Soft constraints categorized to sequence (SE), schedule (SHE) and roster constraints

N	Soft constraint	Category
1.	Maximum number of shift assignments.	SHE
2.	Maximum number of consecutive work days.	SE / SHE
3.	Minimum number of consecutive work days.	SE / SHE
4.	Maximum number of consecutive non-working days.	SHE
5.	Minimum number of consecutive non-working days.	SHE
6.	Maximum number of a certain shift worked.	SHE
7.	Maximum number of consecutive working weekends.	SHE
8.	Maximum number of working weekend in a month.	SHE
9.	Requested days off.	SHE
10.	Requested days on.	SHE
11.	Requested shifts on.	SE
12.	Requested shifts off.	SE
13.	Requested shifts for each weekday.	SHE

Chosen methods use different approach on building the schedules of a roster. Shift sequence based method builds the schedules using shift sequences, simulated annealing - using individual shifts.

2.1. Shift Sequence Based Method

This method consists of two stages: generation of shift sequences and schedule construction according to generated shifts, that are discussed in sections 3.1.1 and 3.1.2. Hard and soft constraints are additionally categorized to sequence, schedule and roster [4]:

i) sequence constraints are applied when constructing shift sequences for each cardiologist;
ii) schedule constraints are applied when combining schedule for each cardiologist;
iii) roster constraints are applied when constructing an overall solution - roster.

Categorized constraints are listed in the Table 4 and Table 5. Last column describes which category of constraints listed above it applies to.

2.1.1. Shift Sequences Construction

In this stage, the shift sequences are constructed for each cardiologist, considering sequence constraints. Shifts sequences are ranked by their penalties for easier retrieval in later stage.

To decrease the complexity, it is possible to limit the number of possible valid shift sequences by either considering only sequences with a penalty below a certain threshold, or by selecting the certain amount of the best sequences for each cardiologist in the second stage of the approach. Shift sequence length is up to 5 shifts. If there is a need for construc-

ting sequences of length greater than 5, such sequences are constructed using combination of sequences of length up to 5 shifts. This combination is performed in the schedule and roster construction stage.

2.1.2. The Construction of Schedules

In the second stage of the approach, schedules for each cardiologist are constructed iteratively, using the shift sequences produced in shift sequences construction stage. Only schedule constraints are under consideration then constructing schedule for cardiologists. Roster constraints are applied then schedule is added to roster.

Basic algorithm of *Shift Sequence* (Algorithm 1, see Table 6) is written using method described in Ref. [4]. It is an adaptive iterative method where cardiologists who received the highest schedule penalties in the last iteration are scheduled first at the current iteration.

Schedule construction process is presented in Algorithm 2, see Table 7. It builds a schedule for the cardiologist based on the partial roster built so far for other cardiologists and returns its penalty to Algorithm 1. The basic idea of this algorithm is to generate a schedule with a low penalty value for the nurse, using low penalty shift sequences. Variable `curr_threshold` points what kind of sequences to use, i.e. if its value is 0, then are used only those sequences that has penalty equal to 0. If no valid assignment can be made for the current cardiologist, the shift sequence with the second lowest penalty is considered and so on. The sequences are assigned for the current cardiologist if the penalty of assigning them is under the current threshold (`curr_threshold`).

During the roster construction, and after a schedule has been generated for the current cardiologist, an improvement method based on an efficient greedy local search is carried out on the partial roster. It simply swaps any pair of shifts between two cardiologists in the partial roster, as long as the swaps satisfy hard constraints and decrease the roster penalty.

After all the schedules have been constructed and a roster has been built, there may still be some shifts for which the coverage is not satisfied. To repair this, a greedy heuristic is used. Each extra shift to be assigned is added to the nurse's schedule whose penalty decreases the most (or increases the least if all worsen) on receiving this shift. After this repair step, the local search is applied once more to improve the quality of the overall roster.

2.2. Simulated Annealing Method

The simulated annealing method is used to solve combinatorial optimization problems. A combinatorial optimization problem is a minimization (maximization) problem consisting of three parts: a set of instances; a finite set of candidate solutions for each instance; and a cost function that assigns to each candidate solution for each instance a positive number called cost. The optimal solution to an instance of a minimization (maximization) problem is the candidate solution having the minimum (maximum) cost.

In the simulated annealing method (Algorithm 3, see Table 8), the cost function to be minimized is identified with the energy of a physical system, and the solution space is identified with the state space. The solution space of the optimization problem is explored by a probabilistic hill climbing search, whose step size is controlled by a parameter T that plays the role of the temperature in a physical system.

By slowly lowering the temperature towards zero according to a properly chosen schedule, one can show that the globally optimal solutions are approached asymptotically.

Table 6. Algorithm 1. `Construct_Roster()`

```
construct and rank the shifts sequences for each cardiologist
iteration = 0
set max no. of iterations (MaxNoIter)
randomly order cardiologists
while (iteration < MaxNoIter)
   for "each cardiologist"  in "ordered list of cardiologists"
        Construct_Schedule(cardiologist, partial_roster)
        greedy local search to improve partial roster
        store the best roster constructed so far
        calculate the penalty for the schedule of "each cardiologist"
        sort the cardiologists by their schedule's penalty in a non-increasing order
increase iteration counter
```

Table 7. Algorithm 2. `Construct_Schedule(cardiologist, partial_roster)`

```
set final threshold (f_threshold)
set current threshold (curr_threshold = 0)
while (curr_threshold <= f_threshold)
    for each sequence in  ranked list for the cardiologist do
        for each day from the first day in the planning period
            assign the sequence's  corresponding shifts based on the partial_roster
    if it does not violate any hard constraints and the penalty <= curr\_threshold
    increase the value of f\_threshold
return schedule
```

Table 8. Algorithm 3. Simulated annealing

```
Build initial roster Current_Roster
Initialize starting temperature T
LOOP
    New_Roster = Neighbour of Current_Roster
    Calculation of Current_Cost
    Calculation of New_Cost
    If (Current_Cost - New_Cost <= 0)
    Then
    Current_Roster = New_Roster
    Else
        If (f > Random(0, 1))
        Then
            Current_Roster = New_Roster
        Else
            Do Nothing
    Decrease Temperature
END LOOP When Stop Criterion Is Met
```

Functional parameter f depending on temperature T was calculated using following equation:

$$f = \exp\left[\frac{Current_{Cost} - New_{Cost}}{T}\right] \qquad (2)$$

Simulated annealing chooses a random move from the neighbourhood - if the move is better than its current position then simulated annealing will always take it. If the move is worse then it will be accepted based on some probability.

Basic algorithm of *Simulated Annealing* is written using method presented in Ref. [10]. Neighborhood rosters where created using the following strategies [11].

Single shift-day. The simplest neighborhood of a schedule includes all the feasible solutions that differ in the position of one scheduled shift.

Overtime - Undertime neighborhood. This neighborhood only considers moving shifts from people with overtime to people with undertime.

Personal requests neighborhood. This neighborhood includes personnel personal requirements like shift on or off, day on or off and etc.

Shuffle neighborhood. Instead of moving duties (as in the simple single shift-day neighborhood), all the duties, which are scheduled in a period from one day to a number of days equal to half the planning period, are switched between the person with the worst schedule and any other person.

3. Simulations.

Experiments there held under same conditions: hardware - Double Core CPU 2.16 GHz, amount of iterations - 500, tests were ran separately 100 times. As we see from Table 9, *Simulated Annealing* method creates better quality rosters in shorter time than *Shift Sequence Based* method.

4. Conclusion

This article presents comparison of two methods used in solving real-world medical staff rostering problem. Test results shows that simulated annealing method is more efficient in matter of roster creation time and roster quality.

Table 9. Test results.
E - Average Execution time (s);
R - Average Roster Quality (rel. un.)

Method	E	R
Shift Sequence Based	250	9655
Simulated Annealing	23	8890

References

1. U. AICKELIN; K. DOWSLAND. An Indirect Genetic Algorithm for a Nurse Scheduling Problem – *Computers and Operations Research* 2004, 31(5), p. 761-778.
2. U. AICKELIN, L. JINGPENG. An Estimation of Distribution Algorithm for Nurse Scheduling – *Annals of Operations Research* 2007, 155(1), p. 289 - 309.
3. U. AICKELIN, Uwe; WHITE, Paul. Building Better Nurse Scheduling Algorithms – *Annals of Operations Research* 2004, 128(1-4), p. 159 - 177.
4. P. BRUCKER, E. BURKE, T. CURTOIS, R. QU, G. VANDEN BERGHE. A shift sequence based approach for nurse scheduling and a new benchmark dataset – *Journal of Heuristics* 2010, 16(4), p. 559 - 573.
5. M. BRUSCO, L. JACOBS. Cost analysis of alternative formulations for personnel scheduling in continuously operating organizations – *European Journal of Operational Research* 1995, 86(2), p. 249 - 261.
6. E. BURKE, T. CURTOIS, G. POST, R. QU, B. VELTMAN. A hybrid heuristic ordering and variable neighbourhood search for the nurse rostering problem – *European Journal of Operational Research* 2008, 188(2), p. 330 - 341.
7. A. IKEGAMI, A. NIWA. A subproblem-centric model and approach to the nurse scheduling problem – *Mathematical programming* 2003, 97(3), p. 517 - 541.
8. L. JINGPENG, U. AICKELIN. Bayesian Optimisation Algorithm for Nurse Scheduling, Scalable Optimization via Probabilistic Modeling – *Springer* 2006, p. 315 - 332. – ISBN 978-3-540-34953-2.
9. M.N. AZAIEZ, S.S. AL SHARIF. A 0-1 goal programming model for nurse scheduling – *Computers and Operations Research* 32, p. 491-507.
10. S. KUNDU, M. MAHATO, B. MAHANTY, S. ARCHARYYA. Comparative Performance of Simulated Annealing and Genetic Algorithm in Solving Nurse Scheduling Problem – *Proceedings of the International MultiConference of Engineers and Computer Scientists* 2008 Vol I, 2008, 96-100 p. – ISBN 978-988-98671-8-8.
11. E. BURKE, P. DE CAUSMAECKER, S. PETROVIC, G.VANDEN BERGHE. Variable Neighbourhood Search for Nurse Rostering Problems – *Book Metaheuristics* 2004, 153-172 p. – ISBN 1-4020-7653-3.

Information Technologies In Investment Operations

Jonas Žaptorius[1] [a] , Narghiza Sulaymonova[2] [b]
[1] Vilnius Gediminas Technical University, Sauletekio ave. 11, LT-10223 Vilnius, Lithuania
[2] Tashkent Finance Institute, Kichik halqa yoli street 7, Tashkent 100060, Uzbekistan

Received 3 May 2012, accepted 14 July 2012

Abstract. As national economies are linked together by the exchange of goods and services and by public and private communications networks, global securities markets develop. In securities markets, the introduction of automation, as well as any serious transformation of the enterprise is a complex and often painful process. But securities trading on a global scale brings with it new risks, as well as beckoning opportunities. Investors and Regulators and policymakers are seeking to understand these risks and appraise the demands that they will place on markets, market participants, and their regulators. This article describes the forces encouraging the development of international securities markets, the obstacles that must be overcome, and the major sources of information technology. It provides some estimates of the present extent of cross-border trading, and describes the largest and most active organized markets competitors in providing securities related services in Central Asia and the European Countries. It also describes the important clearing, settlement, and payment mechanisms that support major markets using different IT methodologies in financial market. Finally, it outlines the questions to be faced how to make span of securities trading stretches beyond the scope of national regulatory regimes with smart automation steps.

Citations: Jonas Žaptorius, Narghiza Sulaymonova. Information Technologies In Investment Operations – *Innovative Infotechnologies for Science, Business and Education,* ISSN 2029-1035 – **1(12)** 2012 – Pp. 7-13.

Keywords: Securities market; Smart automation steps; Back-office; Globalization; 1C-Rarus; Mutual funds; Signator/2000.
Short title: Information technologies.

Introduction

In recent years, several stages of formation and development of securities market have passed. During this period the world has undergone many changes: there were ups, downs, and more recently, a severe crisis. Nevertheless, the market grew and developed. And of course, with the country's market economy development a "sea" of financial information appeared - from the dry figures of different trading platforms to the news that could affect the further development of events in a particular market sector. In this regard, investors of all types are in need for timely and complete information to make good investment decisions.

Two rules formulated by Bill Gates in 2005 claim following statements. The first rule of any technology used in a business is that automation applied to an efficient operation will magnify the efficiency. The second is that automation applied to an inefficient operation will magnify the inefficiency [1].

Along with the development of the stock market, markets of information technology developed actively as well. The large flow of information coming from various sources requires treatment. In this case there is a need for proper analysis of information, how to use it in solving certain problems. After all the information is not an end in itself. Any information should be systematized and analysed. It is especially important to find proper and timely solution for investors who are not professional stock market participants that are not directly connected to the trading systems. The problems faced by these investors are quite traditional. In the current context of the global socio-economic development, particularly important area was the provision of information management process, which consists of collecting and processing information necessary to make informed management decisions [2].

Before the governing body is put to the task of obtaining the information, it's processing, as well as generation and transmission of new information in the form of the derivative control actions. These impacts are carried out in the operational and strategic aspects and based on previously obtained data on the accuracy and completeness of which depends largely on the successful solution of many problems of gover-

[a]Corresponding author, email: *jonas@4team.biz*
[b]Email: *nartiss17@gmail.com*

nance. It should be noted that any decisions require processing large volumes of information; competence manager depends not only on past experience, but on the possession of sufficient information on the rapidly changing situation and the ability to use it. Things you should know and understand the future leaders. There is no doubt that the key to success will be the ability to clearly orient in the flow of information and the ability to effectively use this information.

Computerization in the management of economic processes requires, above all, increasing worker productivity by reducing the cost / production, as well as training and professional competence of specialists engaged in management activities. In developed countries, while two are mutually connected revolution in information technology and business is going on.

2. Information technology in the securities industry and functional flow of financial instruments

One of the few offered by today's new integrated information system is *Signator/2000*. The prototype of this system was the system for the automation of investment funds *SIGNATOR*, developed in West Germany. It was created by *Servo Comp GmbH* - known German developer of application software packages. For example in Russia, the system delivers *Signator/2000 Servo Comp* computer company, representing the interests *Servo Comp GmbH* in the Russian market and dealing with the German adaptation of such systems applied to Russian conditions with the use of modern technologies of *ORACLE*. The company completed improvements *Servo Comp Signator/2000* system taking into account peculiarities of the Russian stock market and banking legislation and ensure its implementation and technical support [3].

The system is also available through distributors *Servo Comp* - known Russian companies, Open Technologies, Technology, and cognitive Stins Coman. *Signator/2000* stock system is universal and can be used to maintain the registry customers, accounting for sales, registrar and depository operations, and organization of internal documentation and reporting. Many organizations within service industries such as government agencies, banking and healthcare decide to structure their business with the back office - front office design; in this setting the back office handles tasks not involving the customer, while front office involves those activities that deal with the customer through some form of contact or receive input from them. When the time comes and an organization wishes to improve the back office area and achieve enhanced efficiency and speed; it is commonly suggested that outsourcing should help introduce the intended gains [4].

However, outsourcing is not always the right option for an organization, depending on the activities the back office performs and the organization's size might not make it a supreme candidate for this. It is at this point that the organizations are left standing in the cold as no alternatives are suggested; therefore creating a push towards outsourcing that might end unsuccessfully. The software product "1C-Rarus: Mutual funds, Revision 2" (1C-Rarus: PIF, red.2) is designed to automate the account open, interval and closed-end mutual investment funds of all kinds, as well as retirement savings. It included the following functions.

1. Accounting for Securities Auto loading issue of securities (NDC).
2. Automatic download securities prices (MICEX, RTS).
3. Flexible configuration of the loader deals in user mode.
4. Automatic evaluation of securities, taking into account the priorities of the exposed exchanges.
5. Ability to automatically download applications and the shareholders of the reporting agent.
6. Ability to automatically load movements on shares, and contact information for shareholders from a report by the registrar.
7. The system alerts the onset of corporate events and direct execution of routine transactions, such as repayment of the ACI, the repayment of the bonds, the partial repayment of the bonds.
8. Perhaps auto control Securities and ACI for the period.
9. A universal mechanism that allows carrying out conversion, consolidation, division, calculating, including complex cases. ACI revaluation surplus, revaluation of securities and other property.
10. Accounting for deposits and interest. Consideration of Bills. Accounting and automatic discounting of bad debts.
11. Accounting and revaluation of foreign currency assets and liabilities [5].

3. The automation system of the investment company

It is especially important to take proper and timely solution for investors who are not professional stock market participants that are not directly connected to the trading systems. The problems faced by these investors are quite traditional. Require immediate and comprehensive information delivered in an easy to use and analyse the form. If possible, it should be a program that could automatically analyse incoming information [6].

In any case, one goal is pursued - to maximize the effective investment of funds. From this basic information and other minor problems may arise: the right financial planning streams of payments, risk management, ensuring the optimal balance between profitability and liquidity of assets and simple automation of business activities of the company, from accounting and finishing operations coupled with the regional offices. The process of Investment Company as a professional participant of the stock market is made up of transactions and their execution (see Fig. 1).

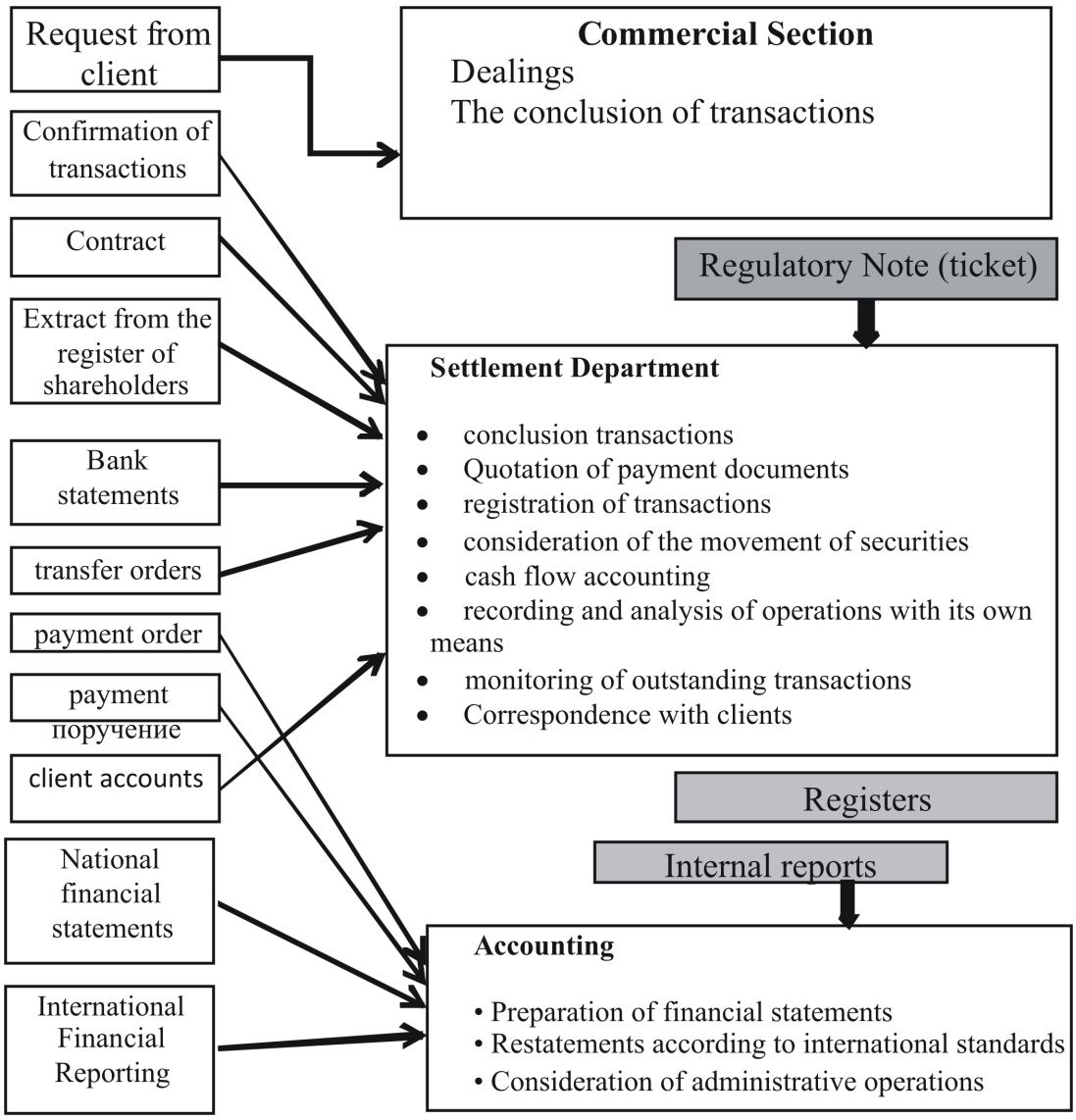


Fig. 1. The process of Investment company (adapted according Ref. [7].)

The investment company may work with a variety of securities and their derivatives on different exchanges. There is no doubt that creativity is the most important human resource of all. Without creativity, there would be no progress, and we would be forever repeating the same patterns (Edward de Bono, 1969). Currently, a large proportion of the turnover of the company has on corporate securities. The largest concentration of deals with them is on OTC trading system. Therefore, we consider the technology works typical of this trading system. In the structure of brokerage (dealer) companies are the following units associated with the process of execution and accounting of transactions (see Fig. 2).

Activity accounting is accounting for transactions directly related to trade in securities. Of course, the right is a statement about the relative autonomy of the back-office and accounting, but we cannot consider this process as a completely unrelated activity, since they reflect the state of the same economic processes, but in different ways. Moreover, the synchronization of business processes of the two units is the key to the organization of concerted action across the company. According to standards developed by NAUFOR, back-office business performs operations corresponding to the execution of transactions, using the traditional system of double entry bookkeeping.

The difference is that the back-office uses a special chart of accounts. The structure chart of accounts except for back-office tracking includes tracking of ownership of securities. This allows you at any time to carry out verification of the location of these securities registrars. Therefore, if kept in sync of the back office and accounting, you can avoid further divergence of balance data with registrars. This is just one of the virtues. The other is that since the two departments are working with the same instruments, albeit in a different perspective, the process of execution and recording of transactions is not only self-regulating, but also more dynamic. And it allows the company to increase sales with the same headcount.

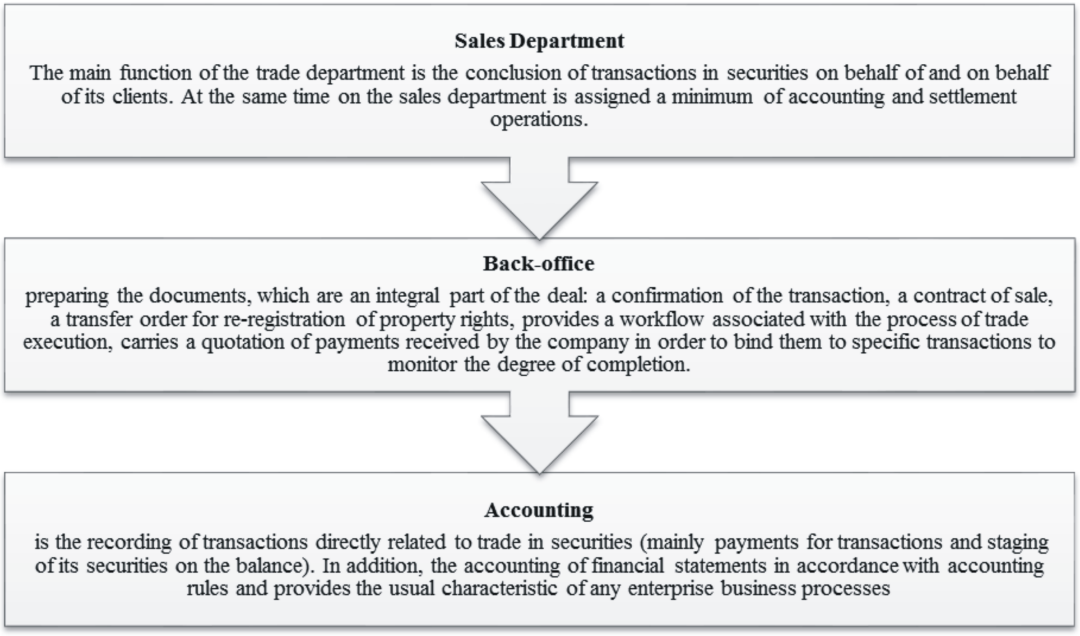


Fig 2. Process of execution and accounting of transactions (adapted according Ref. [8].)

4. Smart automation steps

Automation - a project that should be seen and manageable. Smart Automation - it is not just installing your software that helps automate some parts of the work. This is the process of innovation, which aims to bring business to the next level. "The first rule of any technology used in a business is that automation applied to an efficient operation will magnify the efficiency. The second is that automation applied to an inefficient operation will magnify the inefficiency" (Bill Gates, 2002). Make it a viable, efficient, profitable. This investment the main indicator of which has return on investment. Automation of business - this is one of the elements of modernization as it should not be blindly out of control. As with any opening of a new branch of any marketing project, it should clearly be planned, designed, fit in with the budget. This is the main idea of a phased-development return on investment and its calculation for each of the cycles of work.

In contrast to the typical settings - a phased development is a highly liquid product. It was originally planned, taking into account the specifics. Based on this, start the first stage and it is fully operational and begins to work.

Using your own built-in tools you can immediately carry out the analysis. See how much you spent on this stage, and how much money is brought to you. To carry out the calculation of return on investment. Having the numbers you can make a decision on funding the next phase.

Count: how much additional money earned you want to send for further automation. Or think if it is right to put back the entire increase in development.

Alternatively, return on investment is so effective; you need to have to increase funding. In any case, automation is sighted, managed investments, return on which the well is calculated.

But most importantly it earns on itself. By the time will be put into operation the whole system, it not only pay for itself, but also to bring in excess of the profits.

5. The globalization of IT services: using different methodologies to improve of work staff qualification

When implementing corporate information systems in most cases there is active resistance to field staff, which is a major obstacle to the consultants and is fully capable to prevent or significantly delay the project implementation. This is due to several human factors: the common fear of innovation, conservatism (e.g., storekeeper, has worked 30 years with a paper card file, usually psychologically difficult to change the computer), fear of losing their jobs or lose their indispensability, the fear of substantially increasing responsibility for their actions. The leaders of the company who made the decision to automate their businesses, in such cases should make every effort to promote responsible group of experts conducting the implementation of an information system, to raise awareness of staff [9].

The development of and dependence on technology in the securities industry has accelerated in the last few years, due to increasingly lower margins and the extension into new markets. The ITIO qualification gives an excellent opportunity to the industry community to enhance their knowledge of the role and the challenges of information technology in today's world (Tech Mahindra).

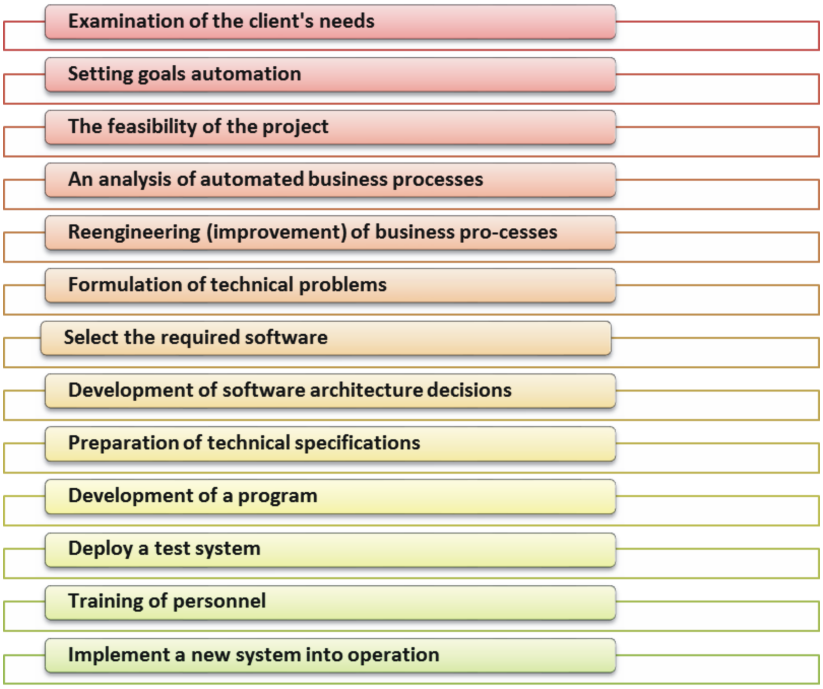


Fig. 3. Smart automation steps (adapted according Ref. [10].)

The Chartered Institute for Securities & Investment is the largest and most widely respected professional body for those who work in the securities and investment industry in the UK and in a growing number of major financial centers round the world. Professionals within the securities and investment industry owe important duties to their clients, the market, the industry and society at large. Where these duties are set out in law, or in regulation, the professional must always comply with the requirements in an open and transparent manner (see Fig.3).

First, one of the most important features of the head of corporate information systems is modules of management accounting and financial controlling. Now, each functional unit can be defined as a centre of financial accounting, with the appropriate level of financial responsibility of its head. This in turn increases the responsibility of each of these leaders, and provides the hands of senior manager's effective tool for the precise control of individual performance plans and budgets. Do not assume that working in the presence of an automated control system will be easier [11].

On the contrary, a significant reduction in red tape accelerates and improves the quality of processing orders, raising the competitiveness and profitability of the enterprise as a whole, and all it requires more discipline, competence and responsibility of the performers.

It is possible that the existing production facilities will not cope with the new flow of orders, and it too will need to make organizational and technological reforms, which subsequently have a positive impact on the prosperity of the enterprise.

A particularly important issue is the selection of the head of the group and the administrator of the system. Head, in addition to basic knowledge of computer technology, must have extensive knowledge of business and management. In practice, in the major Western companies such person has served as CIO (Chief Information Officer) which is usually the second in the hierarchy of management. In domestic practice, the introduction of systems such a role, as a rule, is head of the ACS, or similar to it.

The basic rules of the organization of the working group expresses the following principles [12].

1. Professionals working group should be used with the following requirements: Knowledge of modern computer technology (and the desire to develop them in the future), interpersonal skills, responsibility and discipline.
2. With a special responsibility to approach the selection and appointment of the administrator of the system, since it will be available to nearly all corporate information.
3. The possible dismissal from the group of experts in the process of implementing the project may adversely affect its results. Therefore, group members should be selected from a dedicated and reliable employees and develop a system to support this commitment throughout the project.
4. Once the staff members of the group implementation, the project manager must clearly paint the circle solved each of these tasks, forms and reports of plans, as well as the length of the period. In the best case, the reporting period shall be one day.

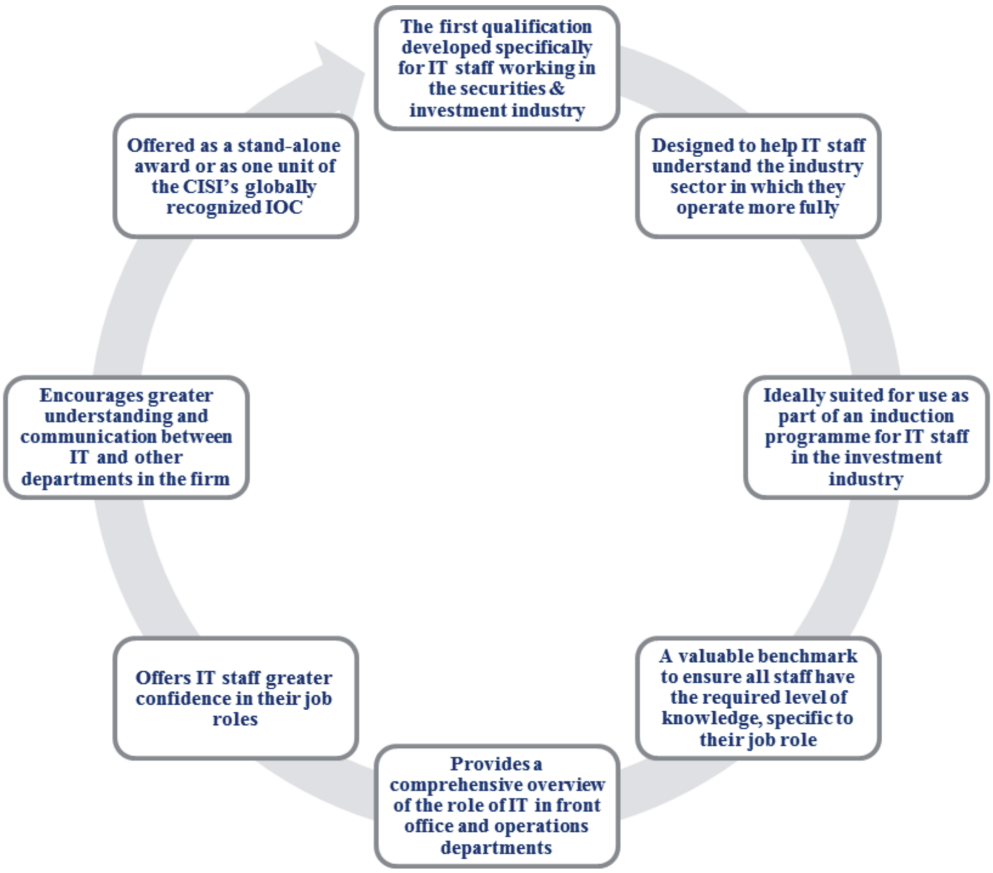


Fig 4. Functions of CISI (adapted according Ref. [5].

6. Conclusion

Of course before starting the automation of business -it does not matter, the average company, large company or a very modest company - we should first of all decide for ourselves what we expect from the program. The fact that life should be better and happier is understandable. Everyone wants work to become easier. But what exactly the company expects from the business automation systems. Which areas require upgrading in the first place, what functions should be run applications. In short, what should be the "ideal program"? Unfortunately, at present, Uzbekistan has not fully formed national approach to financial market, and is currently managing the Uzbek is a volatile mix of Western management theory (which in many respects is not adequate to the current situation) and the Soviet-Russian experience, which, although and largely in harmony with the general principles of life, but it does not meet the stringent requirements of market competition.

The amount of information that must be processed to produce effective management decisions is so large that it has long surpassed human capabilities. It is the modern challenges of managing large-scale production that led to the widespread use of computer technology, the development of automated control systems, which required the creation of new mathematical tools and methods of mathematical economics. By improving information provision, positive results are the following.

1. Possible cost savings by reducing payroll, utilities, cost of software, the cost of mail, the cost of registration of contracts, the costs of redistribution of resources.
2. Elimination of the possible costs in the future: to prevent the future growth of the number of staff, reducing the requirements for data processing, reduction of maintenance costs.
3. The possible intangible benefits: improved quality of information, increased productivity, improved and faster service, new capacity, more confident decisions, improved controls, reduction of late payments, making full use of the software.

The successful development of an integrated financial system is needed that allows solving the following problems: firstly, the production of regular security industry problems and secondly - the problem of choice and order of implementation of information systems. From the principle of unity of information implies the need to eliminate duplication and isolation of its various sources. This means that every economic phenomenon, every economic act shall be recorded only once, and the results can be used in accounting, planning, monitoring and analysis. Thus, the information system should be formed and perfected in the light of the above requirements, which is a necessary condition for improving the efficiency and effectiveness of management.

References

1. Kletzer Lori G.; Robert E. Litan. A Prescription to Relieve Worker Anxiety. – *International Economics Policy Brief* 01-2 (2002) – Washington: Institute for International Economics.
2. Baranowska T.P. Information systems and technologies in economy – *Moscow: Finances and Statistics* 2003. – 416 p.
3. Jorgenson Dale W.; Kevin J. Stiroh. Raising the Speed Limit: US Economic Growth in the Information Age. – Harvard University, 2000. – <http://post.economics.harvard.edu/faculty/jorgenson/papers/dj_ks5.pdf>, accessed 09 04 2012.
4. M. Minaev. Everything rests in Information Technology. – *Securities Market* 2002. N15, p. 47-48.
5. Report and Accounts 2009/10, – *CISI, Chartered Institute for securities and investment* – <http://www.cisi.org/BOOKMARK/WEB9/COMMON/LIBRARY/FILES/ABOUTUS/SHORT2010.PDF>, accessed 09 04 2012.
6. Gaskarov DV. Intelligent Information Systems – *Moscow: High School* 2003. – 432 p.
7. A. G. Ermakov, A tool for achieving System (automation investment company) – *The securities market of 2002* N22 – P.151-155.
8. Improvement of information technology - a necessary condition for the development of the stock market – *Journal of Securities market* 24 (2002) 78-79
9. Mann Catherine L.; Jacob Kirkegaard. Globalization of Information Technology Firms and the Impact on Economic Performance. – Institute for International Economics, 2003 May 2.
10. Sergey Lvov. Smart automation - a technology breakthrough of the crisis. – <http://slvov.ru/content/ articles/umnaja_avtomatizacia.html>, accessed 09 04 2012.
11. Foster Lucia; John Haltiwanger; C.J. Krizan. The Link Between Aggregate and Micro Productivity Growth: Evidence from Retail Trade. – *NBER Working Paper 9120* August (2002). – Cambridge, MA: National Bureau of Economic Research,
12. Kirkegaard Jacob. Stains on the White Collar. – Institute for International Economics, 2003 June).

New Trends In Informatics Study Programmes

Eugenijus Valavičius [a] , Violeta Jadzgevičienė
Vilnius Business College, Kalvariju str. 125, Vilnius, Lithuania

Received 5 June 2012, accepted 11 July 2012

Abstract. Objectives of study program *Media and computer games* are presented and discussed. Online questionnaire was sent to social partners which use media technologies or create computer games. Created primary version of objectives was verified by a survey of relevance.

Citations: Eugenijus Valavičius, Violeta Jadzgevičienė. New Trends In Informatics Study Programmes – *Innovative Infotechnologies for Science, Business and Education,* ISSN 2029-1035 – **1(12)** 2012 – Pp. 14-16.

Keywords: Media and computer games; Informatics study.
Short title: New trends.

Introduction

The demand for Informatics specialists in Europe is questioned last decade at academic and industry level. The European Commission announced the Digital Agenda for Europe two years ago in conjunction with the Europe 2020 Strategy. Digital Agenda is based on 7 pillars: a vibrant digital single market, interoperability and standards, trust and security, fast and ultra-fast internet access, research and innovations, enhancing digital literacy, skills and inclusion, ICT-enabled benefits for EU society.

Over the past 15 years, half of European productivity growth was driven by information and communication technologies, and it is likely that this trend will continue. Almost 40% of productivity and 25% of GDP are related to ICT. According to the Department of Statistics of Lithuania the number of employees in ICT services increased by near 30% during the last five years. Results of study of "Infobalt" confirm that the demand for ICT specialists will grow - by 2016 there will be a need for 21 000 professionals [3].

Lithuania's progress strategy "Lithuania 2030" states that changes must occur in these key areas: smart society, smart economy, smart management. One of the major initiatives of changes in the smart economy is the development of modern information technology and digital infrastructure. The program of development of Lithuanian Information Society for 2011÷2019 has three priorities:

i) the improvement of ICT skills for Lithuanian population;
ii) the development of e-content and e-services, the promotion to use;
iii) the development of ICT infrastructure.

The rapidly expanding fields of application of ICT have an impact on growth in demand for ICT professionals. Such a specialist should be able to handle digital content of ever increasing quality (especially audio and video) and to present it to consumers in understandable way.

1. New study programmes in Informatics

The growth of internet speed and the spread of digital contents encourage the development of more beautiful services based on multimedia. According to European Commission now about two-thirds of mobile data traffic consists of audiovisual content. A significant proportion of the ICT market takes computer games market. A study of European Parliament was initiated in 2009.

One of its finding states that computer games are not harmful to children - on the contrary, playing is healthy. Video games can stimulate developing of strategic thinking, creativity, cooperation and innovative thinking. The study also says that the skills acquired in early age by playing games will remain with us until the end of life.

Of course the researches warn children not to play games intended of adult. Violence and other similar forms of games can have exactly the opposite effect on a child's development. The study also recommends the use of games in schools for learning purposes. There is a public school *Quest to Learn* in the USA which uses computer games not as supplementary but as a main learning tool. Primary school pupils can play computer games such as *Little Big Planet* and *Civilization* as well as the role play and card games.

Situated in Manhattan, school hopes to educate the great mathematicians, inventors, historians, writers and evolutionary biologists. There is also claimed that the school will ensure that their original teaching method will help to achieve great results and adds that it complies with all New York State Education Standards.

[a]Corresponding author, email: *eugvalav@gmail.com*

One of the studies of Ireland Limerick University states that the gaming industry worldwide has surpassed the film industry. Computer game development is an innovative and promising business viewing from a creative and technological perspective. It provides career opportunities for energetic and creative IT students.

Modern students form the generation which was growing in interactive media world from a young age. They have a different way of thinking and operating culture. Taking into account these social and psychological changes in society, educational institutions began to develop new curricula, to introduce new innovative teaching methods in the education process.

There are some similar study programs of technologies in Lithuanian universities and colleges: *Multimedia technologies* in Kaunas Technological University, *Multimedia and computer-aided design* in Vilnius Gediminas Technical University, *Technology of e-publishing* in Alytus College, *Multimedia technology* in Kaunas College. Also there is one similar study programme of informatics: *Applied programming and Multimedia* in College of Social sciences. Vilnius Business college expects to attract students with a new-enough area of computer games which is not taught in other colleges. Also the study programme will introduce to students programming for mobiles and smart devices, programms for robots will be created during training practices. According to the annotations of AIKOS system there are only two analogous university degree programs in Lithuania: Kaunas University of Technology and Vilnius Gediminas Technical University.

2. Objectives of Vilnius Business college's study program *Media and computer games*

The primary version of objectives and competences for the new study program was created in a small workgroup with participation of social partners of the college (Užupis Creative Cluster, Akira Mobile) who have an experience in media and gaming programming. Vilnius Business College has been educating specialists of informatics (Programming, Internet technologies) for a long time so a new study program was created in the same area of studies.

Created primary version of objectives was verified by a survey of relevance. Online questionnaire was prepared in Lithuanian and English. Invitations to answer questions were sent to college's social partners and other companies with priority to companies which use media technologies or create computer games. 28 companies answered questionnaire including 7 foreign partners and IT companies. Some companies sent short comments or advice (Table 1).

66% of respondents are directors of companies or departments, about half of respondents have 10 or more years of work experience. Respondents consider as most important knowledge and skills of computer-aided design, graphics and visualization of information, games programming with Java and specialized tools or libraries (93% answers "Is necessary"), structural and object programming with C, C++ and C#, computer games projecting and creating, collaboration and team work (89% answers "Is necessary"). Two- and three-dimensional graphics, animation, creation of a modern game design using knowledge of human and computer interaction skills were evaluated a little below (above 80% of positive responses).

Respondents considered knowledge of global and local networks, maintenance of computer hardware (93% of answers "Not necessary" or "Has no use"), installing of OS and applications, user consulting as less important (75% "Not necessary" or "Has no use"). Also database knowledge was evaluated as not important (only 54% of answers "Necessary").

Assessments of Lithuanian and foreign partners had a significant difference evaluating knowledge of artificial intelligence and robotics ("Necessary" choose 50% of Lithuanians and 85correct communication in Lithuanian and foreign language ("Necessary" choose 57% of Lithuanians and 100% of foreign respondents), knowledge of audio and video processing ("Necessary" choose 67% of Lithuanians and 100

Respondents also left separate remarks. They offered to pay more attention to knowledge of algorithms and data structures, computer architecture especially those topics that are related to the operation of computer games: memory allocation, video cards, CPU performance. Also the need of mathematical knowledge (linear algebra and geometry, logic, numerical methods), responsibility, ability to reach the goal and to work in team was emphasized. Foreign partners offered to give an additional focus on web technologies (HTML5, PHP, etc.).

Based on these results number of credits for computer networks, operating systems was reduced, separate notices were transferred to teachers who create programs for mathematics, computer architecture, data structures and algorithms, project management subjects. More subjects will use project work as a form of assessment.

After survey and analysis the following goals were set for study program:

a) to use and supervise software and hardware, to apply cloud services;
b) to develop, test and debug applications, databases, using modern programming tools and techniques;
c) to create, test and debug applications and databases using modern programming languages and technologies;
d) to apply internet technologies to develop and provide online services and entertainment;
e) to develop and integrate media products for a wide range of information environments (smart devices, etc.);
f) to design and develop interactive games for computers and smart devices;
g) to work in a team, to assess a value of knowledge and communication between people, nature and technology, to have a professional responsibility.

Table 1. Assessment of knowledge and skills of study program (in %).
1 - necessary; 2 - not necessary; 3 - has no use (I don't know)

N	Knowledge and skills	1	2	3
1.	To apply information technologies (computer, mobile, NFC, GIS, etc.) in solving practical tasks.	71	21	7
2.	To exploit the IT hardware, to install and configure operating systems, application software, to consult customers.	25	61	14
3.	To design, implement and exploit the local computer network, to use the global network services.	7	68	25
4.	To understand the rules of data structures creation and storage and to create new data structures. To analyze and create information-processing algorithms.	68	29	4
5.	To develop, to test and to debug programs using principles of structured and object-oriented programming (C, C++, C #).	89	7	4
6.	To have an understanding of artificial intelligence technology, basics of robotics.	61	36	4
7.	To understand the principles of database creation and management, to create and exploit the databases (mySQL).	54	39	7
8.	To know the World Wide Web technologies (HTML, XML, CSS) and possibilities of their application. To analyze and develop the content management systems.	64	29	7
9.	To create static and dynamic elements of the websites, to apply dynamic principles of the website design (Adobe Photo Shop, Adobe Illustrator, Adobe Flash). To design and program the subsystems websites' security (PHP, etc.).	61	29	11
10.	To have a general theoretical framework of computer design, on which media technologies are based. To apply the principles of computer graphics and information visualization for effective creation of multimedia systems.	93	4	4
11.	To use audio and video processing technologies.	75	25	0
12.	To combine technologies of 2D and 3D graphics and animation for multimedia systems creation.	86	14	0
13.	To know classification of games, design principles and creation techniques.	89	4	7
14.	To program interactive games using effective tools and libraries (Java).	93	4	4
15.	To create a modern game design using knowledge of human and computer interaction.	86	11	4
16.	To be able to analyze problems arising for the users, to help solve them, to understand the responsibility for the taken decisions.	75	21	4
17.	To comply with the principles of cooperation and ethical standards, teamwork and project work.	89	7	4
18.	To communicate correctly in Lithuanian and foreign language.	68	29	4

Conclusion

The rapid expansion of ICT application areas requires specialists who are able to handle the digital content of constantly increasing quality (especially audio and video) and to make it understandable to consumers. A new study program was designed taking into account the social and psychological changes in thinking and performance culture, Lithuanian and foreign partners' experience. Objectives of study program were focused on video and audio technologies as well as conventional programming for computers and smart devices, all this combined with approved Regulation of Informatics field of study.

References

1. Informatikos studijų krypties reglamentas – Lietuvos respublikos švietimo ir mokslo ministro įsakymas. 2007 m. gruodžio 22 d. Nr. ISAK-2580. – <http://www.smm.lt/smt/st_org/docs/st_regl/Informatika%20akt.pdf>, accessed 2012 03 26.
2. Europos kreditų perkėlimo ir kaupimo sistemos (ECTS) nacionalinės koncepcijos parengimas: kreditų harmonizavimas ir mokymosi pasiekimais grindžiamų studijų programų metodikos kūrimas ir diegimas. – <http://www.ects.cr.vu.lt>, accessed 2012 03 26.
3. Plečkaitis A. IRT specialistų poreikis Lietuvoje. Pasiūlos ir paklausos prognozė 2010-2020 – Infobalt. IRT kvalifikacijų ir kompetencijų rinkodaros projektas. Infobalt. 2011.

Zipf and Related Scaling Laws. 2. Literature Overview of Applications in Linguistics

Giedrė Būdienė [a] , Alytis Gruodis
Vilnius Business College, Kalvarijų str. 125, Vilnius, Lithuania

Received 1 February 2012, accepted 25 February 2012

Abstract. The overview of applications of Zipf and related scaling laws in linguistics are presented where the mathematical formulation of task in the framework of one and multi-dimensional distribution takes place. The object of quantitative linguistics represents the natural (western and eastern) as well as artificial languages (programming languages and random texts). Significant part of applications is devoted to the artificial intelligence systems based on zipfian distributions which are useful for cognitive search mechanisms.

Citations: Giedrė Būdienė, Alytis Gruodis. Zipf and Related Scaling Laws. 2. Literature Overview of Applications in Linguistics – *Innovative Infotechnologies for Science, Business and Education,* ISSN 2029-1035 – **1(12)** 2012 – Pp. 17-26.

Keywords: Zipf law; power law; quantitative linguistics; mathematical linguistics.
Short title: ZIPF law - linguistics - 2.

Introduction

Classification and ordering of selected sets constructed using multiple objects represent an unresolved problem in many areas of urban as well as scientific activity. Power law distributions represent statistical behaviour of classification in order to reselect the frequently used items from random occurred ones. Human speech belongs to one of irregular item distribution, and automatized language recognition (OCR, handwriting recognition, spelling correction etc) as well as artificial intelligence (augmentative communication, chat-boots etc) are based on statistical properties of language.

As a subdiscipline of *general linguistics*, the *quantitative linguistics* (or so-called *mathematical linguistics*) studies the quantitative aspects of structure typical for natural language. Static and dynamical approaches (present status and time-domain) allows understanding the changes in language morphology and undercrossing of several languages in certain field. It is obviously that statistical mathematical methods formulates the models, which are applicable by analysing the natural languages. Formulation of language laws allows to extrapolate the generalities of language into affinity group. Specific terms are used in such type modeling.

Linguistic object represents any text in any language where morphological, semantic and lexical rules were used in order to represent the certain idea.

Item represents an linguistic unit (smallest linguistic object). According to the most popular approach, item corresponds to single word (including all word forms). Another approaches allow to use two-words, three-words, also letters, syllables, morphological or semantic constructions etc. Taking more generally, any combination of *lexemes* (so called "base" words or dictionary-entries) extracted from regular or random texts according to certain rule could be treated as an item. For random text generation, the set of any letters of finite amount also could be treated as an item.

Token as the semantic element of programming or natural language could represent an linguistic unit and could be treated as an item.

Corpus represents structured set of texts (usually electronically stored, large amount) devoted for statistical analysis.

Our previous publication [1] was devoted to the overview the applications of Zipf and related scaling laws in economics. This work is aimed to the overview it in linguistics. We have selected about seventy typical references (up to 2011) including several of historic importance. Three approaches of the mentioned problem are presented below.

1. Mathematical formulation of task in the framework of one- and multi-dimensional distribution; description of the object and related laws in quantitative linguistics; description of models for word frequency distributions.
2. Zipfian applications for natural (western and eastern) as well as artificial languages (programming languages and random texts); principles of formation dictionaries.
3. Artificial intelligence systems based on ranked item frequency; cognitive mechanisms including search; language evolution as an informational process.

Quantitative linguistics is empirically based on the results of language statistics through statistics of any linguistic object.

[a]Corresponding author, email: *giedre@kolegija.lt*

1. Mathematical formulation of task

Zip law. George Zipf found the power-law-like word frequency dependence on word rank. The most frequent word appears twice as often as next most popular word, three times as often as 3rd most popular, and so on. So called zipfian distribution relates frequency $f(r)$ of item occurrence in finite corpus to item rank $r(w)$ according to Eq. (1).

$$f(r) = \frac{\alpha}{r^{\gamma}} \tag{1}$$

$$\log f(r) = \alpha - \gamma \cdot \log r \tag{2}$$

In particular case for Zipf distribution, exponent $\gamma \approx 1$. In a logarithmic scale, this dependence represents a straight line, graphical charts are presenten in previous Ref. [1].

For finite corpus of size N, the Zipf coefficient K_{lan} depending on language could be established according to Eq.(3):

$$K_{lan} = N\frac{r(w)}{c(w)} \tag{3}$$

where $c(w)$ - number of selected ranked items w. The simplest case of Zipf law is the famous f^{-1} hyperbolic function. For detailed textbook, see very large study [2] prepared by Saichev et al.

Li in review article [3] devoted to the 100th anniversary of the birth of George Zipf accented the ubiquity of Zipf law. All questions are not answered yet.

1. Is there a rigorous test in fitting real data to Zipf law?
2. In how many forms does Zipf law appear?
3. In which fields are the data sets claiming to exhibit Zipf law?

Heaps law and Herdan law. Harold Heaps discovered distribution of vocabulary size on text length [4]. Number of distinct items (words) $V(n)$ in a part of text containing n items is exponentially proportional to n - so called **Heaps** law, see Eq.(4).

$$V(n) = \alpha \cdot n^{\gamma} \tag{4}$$

With English text corpora, typically $\alpha \in [10 \div 100]$, and $\gamma \in [0.4 \div 0.6]$. Heaps law is asymptotically equivalent to Zipf law concerning the frequencies of individual items (words) within a text.

Gustav Herdan [5] proposed following formulation: the logarithm of vocabulary size divided by the logarithm of text size is a constant smaller than 1 – so called **Herdan** law, see Eq.(5). It is evident that Eq.(5) corresponds to Eq.(4) when α=1.

$$\gamma = \frac{\log V(n)}{\log n} \tag{5}$$

Task of item classification could be treated as a significant part in signalling theory, which examines communication types between individuals. The nature of the Zipf and Heaps laws is not yet clear. According to statements in signalling theory, the language items distributions via power law expressions seem to be specific for natural as well as artificial languages only, otherwise, another signal systems, based on-demand assumption, show stochastic behaviour only [6].

Alfred Lotka states that the number of authors making n contributions is proportional to the n^{-2}. **Lotka law** describes the frequency of publication $V(n)$ by authors n in any given field. It could be treated as one of a variety of Zipf law (γ=2).

$$V(n) = \frac{\alpha}{n^2} \tag{6}$$

Egghe [7] investigates Herdan law and Heaps law from a purely mathematical and informetric point of view. Dependencies according to Lotka law (exponent γ=2) and Zipf law (exponent γ=1) must be treated as expression of boundary conditions by analysing text in linguistics (citations and regular text, respectively).

Bernhardsson et al. [8] analyse text-length dependence of the power-law index of a single book. They have been found that exponent value decreases from 2 to 1 with increasing text length according to extended Heaps law and Zipf law respectively. Authors proposed an idea that the systematic text-length dependence can be described by a meta book concept, which is an abstract representation reflecting the word-frequency structure of a text.

Analysing on any text usually starts from two operations.

1. Calculating of item frequency distribution on rank. In many cases, Zipf or Lotka dependences are expected, for example, see Fig. 1. Exponent $\gamma \in [1 \div 2]$.
2. Calculating of vocabulary size distribution on text size. As usually, Heaps dependence expected, for example, see Fig. 2.

Fig. 3 represents the plot (in log-log coordinates) of ranked word frequency. English corpus was obtained from Wikipedia (data until November 27, 2006) [9]. As expected for English language, the most popular words are "the"(r=1), "of"(r=2), "and"(r=3), also "a", "in", "be", "to", "that" and so on. Actually, Zipf law represents an harmonic series r^{-1}, which describes the real word distribution quite well in first assumption only.

Initial part of dependence in interval AB is claimed as non-zipfian dependence ($\gamma \approx 0.5$). Part AB represents dependence of the words which are morphologically or semantically requested (according to the language construction).

Middle part of dependence in interval BC represents Zipf law (γ=1).

Part CD represents dependencies of rarely used words, so called citation words according to Lotka law (exponent γ=2). Also long tail dependence in interval CE represents Zipf-Mandelbrot law ($\gamma \geq 2$).

These lines correspond to three distinct parameterizations of the Zipf-Mandelbrot distribution.

Models for Word Frequency Distributions. It has taken more than 100 years to discuss the item frequency occurrence in different sciences and linguistic areas such as natural sciences (particle distribution, gene sequence, and earthquakes), economics (market parameters, growth prognosis),

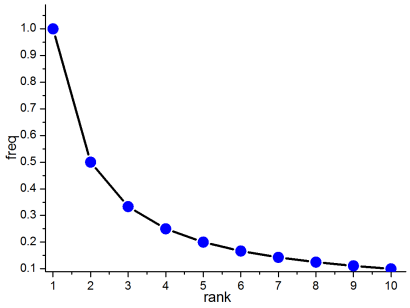


Fig. 1. Zipf law: simulated word frequency distribution on word rank according to Eq.(1), α=1, γ=1.

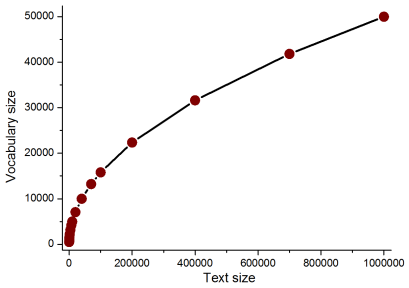


Fig. 2. Heaps law: simulated vocabulary size distribution on text size according to Eq.(4), α=50, γ=0.5.

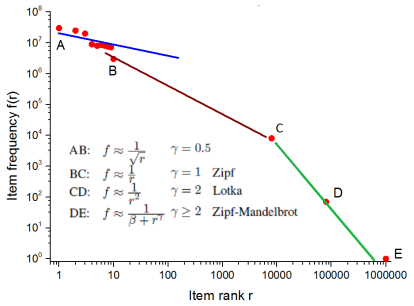


Fig. 3. Ranked word frequency dependence (in log-log scale). English corpus from Wikipedia, Nov 27, 2006. Adapted according to Ref. [9].

urbanistics (the dynamical sizes of cities) etc. Linguistics represents a specific field area in information exchange where chaotic and semi-chaotic sequences – items are encompassed into general power laws. Zipf law uncovers the relationship between word frequency $f(r)$ and its rank r – see Eq. (1).

Zipf [10] states that the *Principle of Least Effort* as the primary principle governs our entire individual and collective behaviour of all sorts, including the behaviour of our language and preconceptions. Kirby [11] analyses validity of Zipfs law using different type of examples. Simulation of organisational behaviour gives the zipfian dependence.

Kosmidis et al. [12] probed to formalize the language system using a simple expression for the Hamiltonian, which is directly implied by the Zipf law. Several language properties such as universality of the Zipf exponent, the vocabulary size of children, the reduced communication abilities of people suffering from schizophrenia could be able to explain.

Historically most important Zipf law according to Eq.(1) and several derived/related laws such as **Paretto distribution** according to Eq.(7) can be applied for strong selection, sorting, prediction, recognition of linguistic items of different languages.

$$f(r) = \left[\frac{r}{r_{min}}\right]^{-\gamma} \tag{7}$$

Egghe [13] analyses graphically the relation between the fraction of the items and the fraction of the sources producing these items. Paretto distribution or so called 80/20-rule by fitting Lorenz curve is evident. Egghe claims that the share of items as a function of the corresponding share of sources increases with increasing size of the system.

Newman [14] reviews some of the empirical evidence for the existence of power-law forms and the theories proposed to explain them. The origin of power-law behaviour has been a topic of debate for more than a century. Darooneh [15] analyses the statistics of ranked words in natural languages using the rank-frequency plot of these words. In that case, model of fractional brownian motion was used in order to improve the power law prediction. Verification of a finite size scaling ansatz was done. This routine allows finding the correct relation between the Zipf exponent and the Hurst exponent characterizing the fractional brownian motion.

Montemurro [16] proposes the revisited **Zipf–Mandelbrot law** in the context of linguistics - see Fig. 4. Its well known that Zipf–Mandelbrot law describes the statistical dependence of the items from certain corpus only. Significant deviations become statistically relevant as larger corpora are considered.

$$f(r) = \frac{\alpha}{(1+\beta r)^{\gamma}} \tag{8}$$

$$f(r) = \frac{\alpha}{\beta + r^{\gamma}} \tag{9}$$

By varying the scale parameter β, it is possible to fit the word frequency dependence in part AB - see Fig.3.

Lognormal law. Fig. 5 represents the probability density function of a log-normal distribution according to Eq.(10). Mentioned non-symmetrical multi-dimensional distribution contains location parameter μ and scale parameter σ.

$$f_X(x,\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \tag{10}$$

Weibul distribution. Fig. 6 represents the probability density function of a Weibull random variable which is useful function in order to simulate particle size distribution. Eq.(11) represent expression for ($r \leq 0$), where positive shape and scale parameters take place: k>0, λ>0.

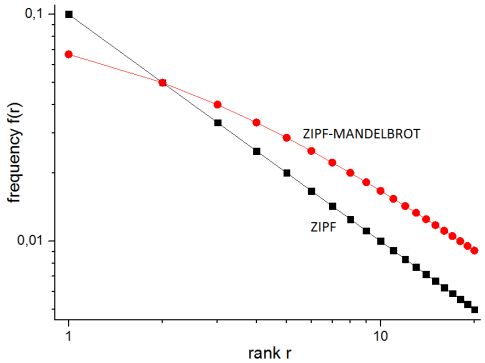


Fig. 4. Simulated word frequency dependence on rank.
Zipf distribution according to Eq.(1), black;
Zipf-Mandelbrot distribution according to Eq.(8), β=0.5, red.
For both curves, α=0.1, γ=1.

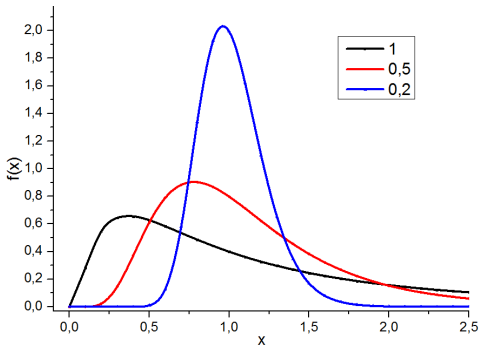


Fig. 5. Log-normal distribution according to Eq.(10), when μ=0. Scale parameter σ=1, black; σ=0.5, red; σ=0.2, blue.

$$f(r,\lambda,k) = \frac{k}{\lambda}\left(\frac{r}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{r}{\lambda}\right)^{k}\right] \qquad (11)$$

Baayen [17] describes three models for word frequency distributions: generalized inverse Gauss-Poisson law; log-normal law according to Eq.(10) and Zipf law according to Eq.(1). Goodness of fit and rationale are commented. It was concluded that no model could vouch the exclusive validity. The role of morphology in shaping word frequency distributions is actual because the vocabulary richness in literary studies correlates to the morphological productivity in linguistics.

Traditionally, zipfian dependencies are devoted for ranked distributions in linguistics, bet otherwise, full statistical analysis such as ANOVA allows analysing the distributions of random and quasi-random items. Limpert et al. [18] analyse applicability of log-normal distribution in several areas of science.

Mitzenmacher [19-20] represent quite brief history of applications related to the several recently proposed models such as lognormal and power law distributions in several areas including validation of models and control of systems. Analysis in many field allow concluding that lognormal distributions have arisen as a possible alternative to power law distributions across many fields.

Menzerath-Altmann law is devoted for certain linguistic construction which contains the constituents: the size of the constituents decrease with increasing size of the construction. Eq.(12) relates $f(r)$ syllable length to r - number of syllables per word. Kohler [21] suggested that linguistic segments contain information about its structure (besides the information that needs to be communicated).

$$f(r) = \frac{\alpha \cdot r^{\beta}}{\exp(\gamma r)} \qquad (12)$$

Aren law (exponential distribution according to Eq.(13)) plays significant role in process modeling. Aren expression could be derived as the special case of Menzerath-Altmann law, when β=0.

$$f(r) = \frac{\alpha}{\exp(\gamma r)} \qquad (13)$$

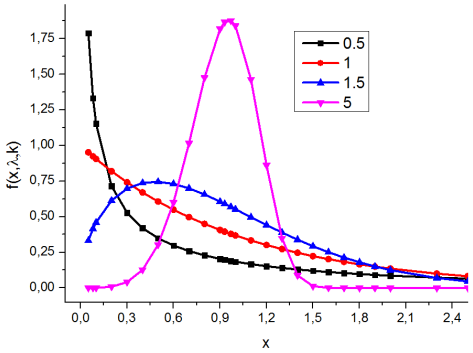


Fig. 6. Weibul distribution according to Eq.(11) with different exponent k={0.5, 1, 1.5, 5} when λ=1.0.

Eliazar et al. [22] presented a universal mechanism for the temporal generation of power-law distributions with arbitrary integer-valued exponents. Hill [23] describes several approaches for applicability of power-law. Generally, deviations from the one-exponential distribution cover the stochastic manifestation of item groups, and for such case data fitting must be done using sophisticated models: **Yule-Simon** distribution according to Eq.(14) or **beta function**, so called Euler integral according to Eq.(15).

$$f(r) = \frac{\alpha \cdot \beta^{r}}{r^{\gamma}} \qquad (14)$$

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t. \qquad (15)$$

Li et al. [24] describe several two-parameter models, including beta function, Yule function, Weibull function for linguistic analysis. Letter frequencies, word-spacing, and word frequencies were used as the ranked linguistic data. Li claims that beta function fits the ranked *letter* frequency distribution the best, but otherwise, Yule function fits the ranked *word-spacing* distribution the best. Altmann, beta, Yule functions all slightly outperform the Zipf power-law function in word ranked- frequency distribution.

Naumis et al. [25] probed to solve the task of fit rank distributions when fits usually fail at the tail. Naumis proposed to use beta-like function. Authors claim that the observed behaviour at the tail seems to be related with the onset of different mechanisms that are dominant at different scales, providing crossovers and finite size effects.

2. Zipfian applications for languages

Dictionaries. Standardization of language always starts from dictionaries made up from corpuses. Formulated from first suggestions as an empirical law, Zipf law represents a universal distribution for natural language not depending on chosen language, word quantity in item, and specific sphere of language usage. Universality of law allows to use it in digital linguistics. Maslov [26] analyses the specific usage of power laws: how linguistic applications could be applied for

dictionary forming. There are several ground principles necessary for resource effectively compiling in order to create the time-saving search. On one hand, dictionaries as specific word sorting form represent a top-application of Zipf law in linguistics; on the other hand, time-domain processes of word item occurrence in sequences cannot be recognized from different slopes of Zipf curves. Willys [27] states that king of ambiguity does not allow us to solve the reverse task. Maslov et al. [28] analysed the big number of formulas of linguistic statistics. The notions of real and virtual cardinality of a sign were introduced. It was concluded that formula refining Zipf law for the occurrence frequencies in frequency dictionaries can be extended to the big semiotic systems. Wyllys [29] provided the jargon stardartization routine in scientific writing using zipfian distributions.

Zanette [30] analyses the family name distribution as time-dependent process. The evolution of family-name distributions is limited vertically depending on cultural features. Empirical power law distribution takes place.

Murtra et al. [31] analyse the applications of Zipf law in the context of a very general class of stochastic systems. Complexity of the description of the system provided by the sequence of observations is the one expected for a system evolving to a stable state between order and disorder. Powers [32] demonstrate how Zipf analysis can be extended to include some of the phenomena not explainable using power-law distributions.

Natural western languages. Regularities of zipfian item distribution are confirmed in many modern languages of great importance, for example American English [33]. Maslov [34] analyses English, French, Spanish languages as the most dominating western languages by expansion among spoken population all over the world. Typical application of Zipf distribution is related to the forming of frequency dictionary. It should be considered that varieties of dictionary might be defined as a logarithmic correction to the Zipf–Mandelbrot law whereas the main problem lies in the tails of distribution. The tails are formed from less-frequently or seldom occurring words derived or constructed morphologically according to the informal rules of spoken and written language.

Tuzzi et al. [35] analyse nonstandard Italian texts such as official presidential speeches in order to confirm the Zipf law. The results showed the unique lexis of the corpus. The analyses allow us to find a position for each president on the synthetism/analytism scale and individual characteristic features of each president.

Popescu et al. [36] presented novel method for language analysing. Even though Zipf law can be applied to a variety linguistic data, a common formula of law cannot be derived to be applicable to the all data sets. New approach to the problem consisting of the multi-component analysis was proposed and tested in 20 languages.

Ha et al. [37] analyse extremely large corpuses: English corpus of 500 million word tokens and 689,000 word types. Is was established the zipfian dependency takes place: the usual slope close to γ=1 for rank less than 5,000, but then for a higher rank it turns to give a slope close to γ=2. Ha concludes that presented phenomenon is done due to foreign words and place names. The Zipf curves for Celtic, Irish languages were presented. Because of the larger number of word types per lemma, it remains flatter than the English curve maintaining a slope of γ=1 until a turning point of about rank 30000.

Ausloos [38] described translation problem: a comparison of two English texts also translated into Esperanto are discussed in order to observe whether natural and artificial languages significantly differ from each other. Word frequencies distribution (studied by a Zipf method) and word lengths distribution (studied by a Grassberger–Procaccia technique) were used. Quantitative statistical differences between the original English text and its Esperanto translation were found. Different power law distributions were observed. The Zipf exponent is equal to $\gamma \in [0.50 \div 0.30]$ depending on how a sentence is defined. Together with the attractor and space dimension, such parameters could also be attached for measurement of the author style versatility.

Programming languages. Zhang [39] discovered the power-law regularities in the distribution of lexical tokens in modern Java, C++ and C programs. It was established that such distributions follow Zipf–Mandelbrot law, and the growth of program vocabulary follows Heaps law.

Natural eastern languages. Natural eastern languages are typical examples of expansive processes of language formation in comparison with western languages. Dahui et al. [40] presented the research where data of traditional and modern Chinese literature was used. Significant differences between Zipf law distributions of mentioned Chinese character sets were found - due to disordered growth of dictionary. Dahui established that the true reason for Zipf law in language is that growth and preferential selection mechanism of word or character in given language.

Ranking problems occur when parallel texts in Chinese and English are analysed according to the frequency distribution. Zipf distribution is applicable until certain barrier of token amount (1 thousand for Chinese and 5 thousand for English). Presence of barrier can be explained by excess of additional tokens, which were put into the context as semantically uncompleted forms. Ha et al. [41] state that when single are combined together with n-gram characters in one list and put in order of frequency, the frequency of tokens in the combined list follows Zipf law - $\gamma \approx 1$. This unexplained behaviour is also found for English 2-byte and 3-byte word fragments.

Xiao [42] analyses applicability of Zipf Law in Chinese word frequency distribution. It was also found out that low frequency words constitute over half of the corpus word occurrences. This is the main reason why data sparse in statistical approaches could not be significantly reduced even ex-

panding corpus scale.

Sen et al. [43] solved the task of validity of Zipf law related to the word (item) length and the frequency was confirmed by analysing the big sets (up to 5,800 words). The main exception is found to be one-letter words.

Changing object of investigation from regular token to specific items – family names – it is necessary to describe the complicated origin of item, which encompasses family name as well as birthplace. Family name distributions with or without the information of the regional origins are applicable to power function - Zipf law. Kim et al. [44] and Miyazima et al. [45] presented the analysis family names belonging to Korean and Japan societies, respectively. In addition, Miyazima states that the relation between size and rank of a family name also shows a power law. Yamada et al. [46] used another fitting technique by means of q-exponential function for the distribution of Japanese family names in order to obey power-law distribution (Zipf law).

Several differences between phonogram-based language (English) and ideogram-based language (Japanese) were found by analysing power law distribution by Nabeshima et al. [47]. It was established that frequency of word usage against rank follows power-law function with exponent γ=1 and, for Japanese ideogram, it follows stretched exponential (Weibull distribution) function.

Sheng et al. [48] analyse the statistical properties of English and Chinese written human language. New approach instead of power law distribution was used: so called *framework of weighted complex networks*. These observations indicate that the two languages may have different linguistic mechanisms and different combinatorial natures. The results display some differences in the structural organizations between the two language networks.

Natural language imitation through random text. Randomly generated texts (RGT) represent sets of items with different probability. Distributions of item frequencies of RGT and English are similar and complies with Zipfs law. Li [49] claims that frequency of occupancy of a word is almost an inverse power law function of its rank and the exponent of this inverse power law is very close to γ=1.

Several methods of text generating could be presented such as *intermittent silence process*. Cancho [50] argued that the real power-law type distribution of word frequencies could be explained by generating a random sequence of characters by means of *intermittent silence process*. According to such method, expected frequency spectrum and the expected vocabulary size as a function of the text size could be efficiently calculated.

Monkey-at-the-typewriter model. Perline [51] describes the application of the classical Mandelbrot *monkey-at-the-typewriter* model as the model where Zipf inverse power law is applicable. An explicit asymptotic formula for the slope of the log-linear rank-size law in the upper tail of this distribution is also obtained. By usage of the same *monkey-at-the-typewriter* model, Conrad et al. [52] showed so called recent confusion, where the rank-frequency distribution follows a lognormal distribution. This special model arises in particular case, where letters are hit with unequal probability.

On the other hand, Cancho [53] demonstrate by means of three different statistical tests that ranks derived from random texts and ranks derived from real texts are statistically inconsistent. Cancho concludes that the good fit of random texts to real Zipf law-like rank distributions has not yet been established.

3. Artificial intelligence systems

Cognitive mechanisms including search. Serrano et al. [54] studied the written text problem in the context of text recognition tasks. Two approaches were used for modeling: Zipfs law and Heaps law. It was established the significant relation between the burst nature of rare words and the topical organization of texts. The dynamic word ranking and memory across documents – such two factors could be treated as a key mechanisms explaining the non trivial organization of written text.

Wyllys [55] analyses implications of Zipf law for the design of information systems. He claims that only vocabulary control could be done using Zipf law. Wyllys says that sentence about universality of Zipf law (that different subject-fields may be characterized by different slopes of Zipf curves) seems to have no practical applications in information system design at present (may be in future).

Blanchard [56] solves the problem of a document retrievals in patent mapping tools. Previous stopword list technique was used – as a system which modified the retrieval words into more powerful (i.e. they dramatically impacts the final output and analysis). Stopword lists depend on the document corpus analysed according to power-law.

Calderon et al. [57] analyse the distribution of words in Spanish texts of Latin-American writers from Zipf law perspective. New approach to Zipf law dependencies was used: the frequency of repetition of a particular word among other different words was analysed in order to solve the linguistic problem using statistical approach.

Kello et al. [58] analyse linguistic activities using scaling laws which suggest the existence of patterns that are repeated across scales of analysis. Variable can vary in region between several types. In that case recurrence of scaling laws has prompted a search for unifying principles. In language systems, scaling laws can reflect adaptive processes of various types and are often linked to complex systems near critical points. Findings of scaling laws in cognitive science are indicative of scaling invariance in cognitive mechanisms.

Caron et al. [59] analyse semantic extraction of word groups belonging to the different regions of interest. Zipf law and inverse Zipf law were used in order to characterize the structural complexity of image textures. The distribution

of pattern frequency was modeled as power law distributions. Method allows the detection of regions of interest, which are consistent with human perception, where inverse Zipf law is particularly significant.

Altmann et al. [60] analyse big corpuses where the language has different levels of formality. These distributions are well characterized by a stretched exponential (Weibull) scaling. Distributions of distances between successive occurrences of the same word display some deviations from a Poisson process. The extent of this deviation depends strongly on semantic type. A generative model of this behaviour that fully determines the dynamics of word usage was developed.

Automatic text analysis is grounded on Luhn assumption [61] that frequency data can be used to extract words and sentences in order to represent a certain document. Losee [62] analyses regularities in the statistical information provided by natural language terms about neighbouring terms. We find that when phrase rank increases, moving from common to less common phrases, the value of the expected mutual information measure (EMIM) between the terms regularly decreases. Luhn model suggests that mid-range terms are the best index terms and relevance discriminators. Interpretation of Zipf law from information theoretic point view was provided. Using the regularity noted above, we suggest that Zipf law is a consequence of the statistical dependencies that exist between terms, described here using information theoretic concepts.

New teaching/learning methods. Vousden [63] uses application of Zipf law in order to choose the English teaching material as spelling-to-sound units. In that case, the quantity and adaptability could be rationalized in high degree. Alexander et al. [64] use application of Zipf law for helping the students to create the interconnection between mathematics and other disciplines.

Language evolution as an informational process. In quantitative linguistics, **Piotrowski** law [65] describes the process of language change through several parameters:

i) vocabulary growth;
ii) the dispersion of foreign or loan words;
iii) changes in the inflectional system etc.

Initial hypothesis (everything in language changes as a result of interaction between old forms and new forms) could be formulized through differential equation:

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = k_t \cdot p_t \cdot (C - p_t) \tag{16}$$

where $\mathrm{d}p_t$ - change in the proportion; p_t - proportion of new forms; k_t - time-dependent function.

Most important solution of mentioned differential equation is presented below. In case, if C=1 and k_t=b, solution represents so-called **logistic** curve for modeling the growth phenomena (α is the integration constant). Fig. 7 represents the

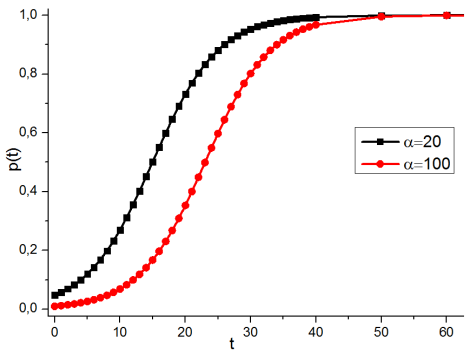


Fig. 7. Logistic distributions for growth modeling according to Eq.(17). β=0.2.
Different integration constant $\alpha \in \{20; 100\}$.

logistic curve made up by Eq.(17).

$$p(t) = \frac{1}{1 + \alpha \exp(-\beta t)} \tag{17}$$

Joshua et al. [66] use mathematical models to analyse the major transitions in language evolution. Word-formation is described as a process related to Shannon noisy coding theorem. Model of the population dynamics of words and the adaptive emergence of syntax is present.

Bernhardsson et al. [67] analyse functional form of the word-frequency distribution. So called null model was used where the words are randomly distributed throughout the text. Initial assumption of sharing characteristics (real novel shares many characteristic features with a null model) was used together with second (functional form of the word-frequency distribution of a novel depends on the length of the text in the same way as the null model). This means that an approximate power-law tail will have an exponent which changes with the size of the text-section which is analysed. The size-transformation of a novel is found to be well described by a specific Random Book Transformation.

Shenon entropy in evolution model. Maillart et al. [68] studied the evolution processes of open source software projects in Linux distributions, which offer a remarkable example of a growing complex of self-organizing adaptive system. The ingredients of stochastic growth models were established empirically which are previously conjectured to be at the origin of Zipf law.

Unpredictability of information content could be characterized by Shannon entropy H where $P(x)$ is the probability that variable X occupies the state x. Summation must be provided over all states N.

$$H(X) = -\sum_{i=1}^{N} P(x_i) \cdot \log_2(P(x_i)) \tag{18}$$

Dover [69] proposed novel formalism of maximum principle of Shannon entropy in order to derive the general power law distribution function. There are big number of examples where Boltzmann entropy is related to the paradigm of

"internal order": complexion, self-interacting, self-organized system etc. Evolution of structure could be modeled by describing the noninteracting conditions since the Shannon entropy is equivalent to the Boltzmann entropy under equilibrium. This formalism was demonstrated in toy model where Zipf law comes out as a natural special point of the model.

Nesterova [70] presented large review of applications of Shenon entropy. Main paradigms - system, structure, information - and corresponding parameters - entropy, negentropy - are desribed for characterization two different - metric as well as information system.

Cancho [71] describes a general communication model where objects map to signals, a power function for the distribution of signal frequencies is derived. Cancho claims that many systems in nature use non-trivial strategies for easing the interpretation of a signal. Presented model relies on the satisfaction of the receiver communicative needs when the entropy of the number of objects per signal is maximized. Estimation in linguistic context is surprising: present exponent ($\gamma\approx2$) is clearly different from the typical of Zipf law ($\gamma\approx1$). It means that Zipf law reflects some sort of optimization. On other hand, the words are used according to the objects (e.g. meanings) they are linked to (linguistic approach).

Cancho [72] analyses the new model for Zipf law proposed for the human word distribution in the framework of information theory: from a no communication phase to a perfect communication phase. Scaling consistent with Zipf law is found in the boundary between phases. The exponents are consistent with minimizing the entropy of words. Presented model is especially suitable for the speech of schizophrenics. Zipf exponent predicted for the frequency versus rank distribution is in a range where γ>1, which may explain the word frequency distribution of some schizophrenics and some children, with $\gamma\in[1.5\div1.6]$. Among the many models for Zipf law, none explains Zipf law for that particular range of exponents. In particular, two simplistic models fail to explain that particular range of exponents: intermittent silence and Simon model.

Conclusion

1. Many linguistic ranked item frequency distributions could be described using Zipf or Zipf-Mandelbrot law with exponent $\gamma\approx1$. Increasing of exponent up to $\gamma\approx2$ (long tail problem) is related to the stochastic nature of items.
2. Yule, beta and Manzerath-Altman distributions could be treated as the "modifications" of more general power-low where specific fitting parameters are useful for precisely adequacy to original distribution.
3. In linguistics, power-law represents influence of human behaviour where language as a communication tool can be used. Dependencies according to Lotka law (exponent γ=2) and Zipf law (exponent γ=1) must be treated as expression of the boundary conditions by analysing text in linguistics.

References

1. Artūras Einikis, Giedrė Būdienė, Alytis Gruodis. Zipf and Related Scaling Laws. 1. Literature Overview of Applications in Economics. – *Innovative Infotechnologies for Science, Business and Education* ISSN 2029-1035 – 2(11) (2011) 27-36.
2. Alexander I. Saichev, Yannick Malevergne, Didier Sornette. Theory of Zipf's Law and Beyond. – Berlin: Springer, 2010.
3. Wentian Li. Zipf's Law Everywhere. – *Glottometrics* 5 (2002) 14-21.
4. Heaps Harold Stanley (1978), Information Retrieval: Computational and Theoretical Aspects, Academic Press. – Heaps law is proposed in Section 7.5 (pp. 206–208).
5. Leo Egghe. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. – *Journal of the American Society for Information Science and Technology* 58(5) (2007) 702–709.
6. Rebecca Bliege Bird, Eric Alden Smith. Signaling Theory, Strategic Interaction, and Symbolic Capital. – *Current Anthropology* 46(2) (2005) 221-248.
7. Leo Egghe. Untangling Herdan's Law and Heaps' Law: Mathematical and Informetric Arguments. – *Journal of the American society for information science and technology* 58(5) (2007) 702–709.
8. Sebastian Bernhardsson, Luis Enrique Correa da Rocha, Petter Minnhagen. The meta book and size-dependent properties of written language. – *New Journal of Physics* 11 (2009) 123015.
9. <http://en.wikipedia.org/wiki/Zipf%27s_law>, accessed 2012.01.15.
10. George Kingsley Zipf. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology – Addison-Wesley Press Inc., 1949.
11. Geoff Kirby. Zipf's law. – *UK Journal of Naval Science* 10(3) (1985) 180-185.
12. Kosmas Kosmidis, Alkiviadis Kalampokis, Panos Argyrakis. Statistical mechanical approach to human language. – *Physica A* 366 (2006) 495–502.
13. L. Egghe. The dependence of the height of a Lorenz curve of a Zipf function on the size of the system. – *Mathematical and Computer Modelling* 43 (2006) 870–879.
14. M.E.J. Newman. Power laws, Pareto distributions and Zipf's law. – *Contemporary Physics* 46(5) (2005) 323-351.
15. A.H. Darooneh, B. Rahmani. Finite size correction for fixed word length Zipf analysis. – *Eur. Phys. J. B* 70 (2009) 287–291 .
16. Marcelo A. Montemurro. Beyond the Zipf–Mandelbrot law in quantitative linguistics. – *Physica A* 300 (2001) 567–578.

17. Harald Baayen. A Linguistic Evaluation. – *Computers and the Humanities* 26 (1993) 347-363.
18. Eckhard Limpert, Werner A. Stahel, Markus Abbt. Log-normal Distributions across the Sciences: Keys and Clues. – *BioScience* 51(5) (2001) 341-352.
19. Michael Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. – *Internet Mathematics* 1(2) (2003) 226-251.
20. Michael Mitzenmacher. The Future of Power Law Research. – *Internet Mathematics* 2(4) (2005) 525-534.
21. Reinhard Köhler. Zur Interpretation des Menzerathschen Gesetzes. – *Glottometrika* 6 (1984) 177–183.
22. Iddo Eliazar, Joseph Klafter. Temporal generation of power-law distributions: A universal 'oligarchy mechanism'. – *Physica A* 377 (2007) 53–57.
23. Bruce M. Hill. The Rank-Frequency form of Zipfs law. – *Journal of the American Statistical Association* 69(384) (1974) 1017-1026.
24. Wentian Li, Pedro Miramontes and Germinal Cocho. Fitting Ranked Linguistic Data with Two-Parameter Functions. – *Entropy* 12 (2010) 1743-1764.
25. G.G. Naumis, G. Cocho. Tail universalities in rank distributions as an algebraic problem: The beta-like function. – *Physica A* 387 (2008) 84–96.
26. V. P. Maslov. Quantum Linguistic Statistics. – *Russian Journal of Mathematical Physics* 13(3) (2006) 315–325.
27. Ronald E. Wyllys. Empirical and Theoretical Bases of Zipf's Law. – *Library Trends* 30(1) (1981) 53-64.
28. V. P. Maslov, T. V. Maslova. On Zipf's Law and Rank Distributions in Linguistics and Semiotics. – *Mathematical Notes* 80(5) (2006) 679–691. – Translated from *Matematicheskie Zametki* 80(5) (2006) 718–732.
29. Ronal Eugene Wyllys. The measurement of jargon standardization in scientific writing using rank-frequency ("Zipf") curves. PhD thesis. – University of Wisconsin, 1974.
30. Damisan H. Zanette, Susanna C. Manrubia. Vertical transmission of culture and the distribution of family names. – *Physica A* 295 (2001) 1–8.
31. Bernat Corominas-Murtra, Ricard V. Sole. Universality of Zipf's law. – *Physical Review E* 82 (2010) 011102.
32. David M. W. Powers. Applications and Explanations of Zipfs Law. – In: D. M. W. Powers (ed.) NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning – ACL, 1998. – Pp. 151-160.
33. H. Kučera, W. N. Francis. Computational Analysis of Present-Day American English. – Brown University, 1967.
34. V. P. Maslov. The Lack-of-Preference Law and the Corresponding Distributions in Frequency Probability Theory. – *Mathematical Notes* 80(2) (2006) 214–223. – Translated from *Matematicheskie Zametki* 80(2) (2006) 220–230.
35. Arjuna Tuzzi, Ioan-Iovitz Popescu, Gabriel Altmann. Zipf's Laws in Italian Texts. – *Journal of Quantitative Linguistics* 16(4) (2009) 354–367.
36. Ioan-Iovitz Popescu, Gabriel Altmann, Reinhard Köhler. Zipf's law—another view. – *Qual. Quant.* 44 (2010) 713-731.
37. Le Quan Ha, Francis J Smith. Zipf and Type-Token rules for the English and Irish languages. – MIDL, Paris, 29-30 novembre 2004. – Pp. 65-70.
38. M. Ausloos. Equilibrium and dynamic methods when comparing an English text and its Esperanto translation. – *Physica A* 387 (2008) 6411–6420.
39. Hongyu Zhang. Discovering power laws in computer programs. – *Information Processing and Management* 45 (2009) 477–483.
40. Wang Dahui, Li Menghui, Di Zengru. True reason for Zipf's law in language. – *Physica A* 358 (2005) 545–550.
41. Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming and F. J. Smith. Extension of Zipf's Law to Word and Character N-grams for English and Chinese. – *Computational Linguistics and Chinese Language Processing* 8(1) (2003) 77-102.
42. Hang Xiao. On the Applicability of Zipf's Law in Chinese Word Frequency Distribution. – *Journal of Chinese Language and Computing* 18(1) (2008) 33-46.
43. B.K. Sen, Khong Wye Keen, Lee Soo Hoon, Lim Bee Ling, Mohd Rafae Abdullah, Ting Chang Nguan, Wee Siu Hiang. Zipf's law and writings on LIS. – *Malaysian Journal of Library & Information Science* 3(2) (1998) 93-98.
44. Beom Jun Kim, Sung Min Park. Distribution of Korean family names. – *Physica A* 347 (2005) 683–694.
45. Sasuke Miyazima, Youngki Lee, Tomomasa Nagamine, Hiroaki Miyajima. Power-law distribution of family names in Japanese societies. – *Physica A* 278 (2000) 282-288.
46. Hiroaki S. Yamada, Kazumoto Iguchi. q-exponential fitting for distributions of family names. – *Physica A* 387 (2008) 1628–1636.
47. Terutaka Nabeshima, Yukio-Pegio Gunji. Zipf's law in phonograms and Weibull distribution in ideograms: comparison of English with Japanese. – *BioSystems* 73 (2004) 131–139.
48. Long Sheng, Chunguang Li. English and Chinese languages as weighted complex networks. – *Physica A* 388 (2009) 2561-2570.
49. Wentian Li. Random Texts Exhibit Zipfs-Law-Like Word - Frequency Distribution. – *IEEE Transactions on information theory* 38(6) (1992) 1842-1845.
50. Ramon Ferrer-i-Cancho, Ricard Gavalda. The Frequency Spectrum of Finite Samples from the Intermittent Silence Process. – *Journal of the American Society for Information Science and Technology* 60(4) (2009) 837–843.
51. Richard Perline. Zipf's law, the central limit theorem, and the random division of the unit interval. – *Physical Review E* 54(1) (1996) 220-223.
52. Brian Conrad and Michael Mitzenmacher. Power Laws for Monkeys Typing Randomly: The Case of Unequal Probabilities. – *IEEE Transactions on information theory* 50(7) (2004) 1403.

53. Ramon Ferrer-i-Cancho, Brita Elvevag. Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution. – *PLoS ONE* 5(3) (2010) e9411. – <www.plosone.org>, accessed 2011.11.19.
54. M. Angeles Serrano, Alessandro Flammini, Filippo Menczer. Modeling Statistical Properties of Written Text. – *PLoS ONE* 4(4) (2009) e5372. – <www.plosone.org>, accessed 2011.11.19.
55. Ronald E. Wyllys. Empirical and Theoretical Bases of Zipf's Law. – *Library Trends* 30(1) (1981) 53-64.
56. Antoine Blanchard. Understanding and customizing stopword lists for enhanced patent mapping. – *World Patent Information* 29 (2007) 308–316.
57. F. Calderon, S. Curilef and M. L. Ladron de Guevara. Probability distribution in a quantitative linguistic problem. – *Brazilian Journal of Physics* 39(2A) (2009) 500-502.
58. Christopher T. Kello, Gordon D.A. Brown, Ramon Ferrer-i-Cancho, John G. Holden, Klaus Linkenkaer-Hansen, Theo Rhodes and Guy C. Van Orden. Scaling laws in cognitive sciences. – *Trends in Cognitive Sciences* 14(5) (2010) 223-232.
59. Y. Caron, P. Makris, N. Vincent. Use of power lawmodels in detecting region of interest. – *Pattern Recognition* 40 (2007) 2521-2529.
60. Eduardo G. Altmann, Janet B. Pierrehumbert, Adilson E. Motter. Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. – *PLoS ONE* 4(11) (2009) e7678. – <www.plosone.org>, accessed 2011.11.19.
61. Luhn H.P. The automatic creation of literature abstracts. – *IBM Journal of Research and Development* 2 (1958) 159-165.
62. Robert M. Losee. Term Dependence: A Basis for Luhn and Zipf Models. – *Journal of the American Society for Information Science and Technology* 52(12) (2001) 1019-1025.
63. Janet I. Vousden. Units of English Spelling-to-Sound Mapping: A Rational Approach to Reading Instruction. – *Appl. Cognit. Psychol.* 22 (2008) 247–272.
64. Linda Alexander, Roger Johnson and John Weiss. Exploring Zipf's Law. – *Teaching Mathematics and its applications* 17(4) (1998).
65. Altmann G., v. Buttlar H., Rott W., Strauβ U. A law of change in language. – In: Brainerd B. (ed.) Historical linguistics. – Bochum: Brockmeyer, 1983. – P. 104-115.
66. Joshua B. Plotkin and Martin A. Nowak. Major Transitions in Language Evolution. – *Entropy* 3 (2001) 227–246.
67. Sebastian Bernhardsson, Luis Enrique Correa da Rocha, Petter Minnhagen. Size-dependent word frequencies and translational invariance of books. – *Physica A* 389 (2010) 330-341.
68. T. Maillart, D. Sornette, S. Spaeth, and G. von Krogh. Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution. – *Physical Review Letters* 101 (2008) 218701 .
69. Yaniv Dover. A short account of a connection of power laws to the information entropy. – *Physica A* 334 (2004) 591-599.
70. Jelena Nesterova. Spatial self-arrangement of expanding structures. 1. Overview of assessment concepts. – *Innovative Infotechnologies for Science, Business and Education* ISSN 2029-1035 – 2(9) (2010) 17-22.
71. Ramon Ferrer i Cancho. Decoding least effort and scaling in signal frequency distributions. – *Physica A* 345 (2005) 275–284.
72. R. Ferrer i Cancho. Zipf's law from a communicative phase transition. – *The European Physical Journal B* 47 (2005) 449–457.