

Natalya Shakhovska
Mykola O. Medykovskyy *Editors*

Advances in Intelligent Systems and Computing V

Selected Papers from the International
Conference on Computer Science and
Information Technologies, CSIT 2020,
September 23–26, 2020, Zbarazh,
Ukraine

Advances in Intelligent Systems and Computing

Volume 1293

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,
Warsaw, Poland

Advisory Editors

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

Rafael Bello Perez, Faculty of Mathematics, Physics and Computing,
Universidad Central de Las Villas, Santa Clara, Cuba

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

Hani Hagras, School of Computer Science and Electronic Engineering,
University of Essex, Colchester, UK

László T. Kóczy, Department of Automation, Széchenyi István University,
Gyor, Hungary


Vladik Kreinovich, Department of Computer Science, University of Texas
at El Paso, El Paso, TX, USA

Chin-Teng Lin, Department of Electrical Engineering, National Chiao
Tung University, Hsinchu, Taiwan

Jie Lu, Faculty of Engineering and Information Technology,
University of Technology Sydney, Sydney, NSW, Australia

Patricia Melin, Graduate Program of Computer Science, Tijuana Institute
of Technology, Tijuana, Mexico

Nadia Nedjah, Department of Electronics Engineering, University of Rio de Janeiro,
Rio de Janeiro, Brazil

Ngoc Thanh Nguyen , Faculty of Computer Science and Management,
Wrocław University of Technology, Wrocław, Poland

Jun Wang, Department of Mechanical and Automation Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Indexed by SCOPUS, DBLP, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST), SCImago.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at <http://www.springer.com/series/11156>

Natalya Shakhovska · Mykola O. Medykovskyy
Editors

Advances in Intelligent Systems and Computing V

Selected Papers from the International
Conference on Computer Science
and Information Technologies, CSIT 2020,
September 23–26, 2020, Zbarazh, Ukraine

Editors

Natalya Shakhovska
Department of Artificial Intelligence
Lviv Polytechnic National University
Lviv, Ukraine

Mykola O. Medykovskyy
Institute of Computer Science
and Information Technologies
Lviv Polytechnic National University
Lviv, Ukraine

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-3-030-63269-4

ISBN 978-3-030-63270-0 (eBook)

<https://doi.org/10.1007/978-3-030-63270-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The book reports on new theories and applications in the field of intelligent systems and computing. It covers computational and artificial intelligence methods, as well as advances in computer vision, current issue in big data and cloud computing, computation linguistics, cyber-physical systems as well as topics in intelligent information management. The papers present new trends in applied linguistics, IT systems, highlight theories and applications of intelligent systems, and discuss big data and cloud computing issues and challenges.

Written by active researchers, the different chapters are based on contributions presented at the workshop in intelligent systems and computing (ISC), held during CSIT 2020, September 23–26, and jointly organized by the Lviv Polytechnic National University, Ukraine, the Kharkiv National University of Radio Electronics, Ukraine, and the Technical University of Lodz, Poland, under patronage of Ministry of Education and Science of Ukraine. All in all, the book provides academics and professionals with extensive information and a timely snapshot of the field of intelligent systems, and it is expected to foster new discussions and collaborations among different groups.

Contents

Applied Linguistics

Experimental Investigation of Significant Keywords Search in Ukrainian Content	3
Oleg Bisikalo, Victoria Vysotska, Vasyl Lytvyn, Oksana Brodyak, Svitlana Vyshemyrska, and Yuriy Rozov	
Linguistic Analysis of Results of Variable Courses Selection by HEI's Students	30
Pavlo Zhezhnych and Anna Shilinh	
Poetic Discourse Processing Applying Graph Theoretic Rules	41
Olena Flys and Maria Bekhta-Hamanchuk	
Textual Features and Semantic Analysis of the Reddit News Posts	53
Solomiia Albota	
A Model of the Information System of the Associative Verbal Network Presentation	71
Olena Levchenko, Oleh Tyshchenko, Marianna Dilai, and Lukas Gajarsky	

Artificial Intelilgence

Intelligent Neural Network Sensory System for the Analysis of Volatile Compounds in Beverages	87
Taras Chaikivskyi, Bohdan Sus, Oleksandr Bauzha, and Sergiy Zagorodnyuk	
AI System in Monitoring of Emotional State of a Student with Autism	102
Vasyl Andrunyk and Olesia Yaloveha	
Neural-like Real-Time Data Protection and Transmission System	116
Ivan Tsmots, Vasyl Rabyk, Oleksa Skorokhoda, and Yurii Tsymbal	

An Analysis of Gene Regulatory Network Topology Using Results of DNA Microchip Experiments	130
Sergii Babichev, Orest Khamula, Iryna Perova, and Bohdan Durnyak	
Dynamic Bayesian Network Model of a Country's Economic Extension	145
Mariia Voronenko, Dmytro Nikytenko, Jan Krejci, Nataliia Krugla, Oleksandr Naumov, Nataliia Savina, and Volodymyr Lytvynenko	
Classification of Objects Based on a Tree-Shaped Artificial Immune Network Model	160
Mykola Korablyov, Oleksandr Fomichov, and Natalia Axak	
Cytological Images Clustering	173
Oleh Berezsky, Oleh Pitsun, Lesia Dubchak, Kateryna Berezka, Taras Dolynyuk, and Bohdan Derish	
Metric Methods in Computer Vision and Pattern Recognition	188
Oleh Berezsky and Mykhailo Zarichnyi	
Dynamic Bayesian Networks Application for Economy Competitiveness Situational Modelling	210
Mariia Voronenko, Dmytro Nikytenko, Jan Krejci, Nataliia Krugla, Oleksandr Naumov, Nataliia Savina, Elzara Topalova, Viktoriia Filippova, and Volodymyr Lytvynenko	
Synthesis of the Centered Bithreshold Neural Network Classifier	225
Vladyslav Kotsovsky, Fedir Geche, and Anatoliy Batyuk	
The Novel Approach to Modeling the Spread of Viral Infections	240
Nataliya Shakhovska, Nataliia Melnykova, Volodymyr Melnykov, Vitaly Mahlovanyj, and Nataliya Hrabovska	
Hybrid Power Plant Control System Based on Machine Learning Methods	251
Aleksandr Gozhyj, Peter Bidyuk, Yoshio Matsuki, Vladyslav Nechakhin, Irina Kalinina, and Oleg Shchesiuk	
Intelligent Systems Based on Ontology Representation Transformations	263
Yevhen Burov, Khrystyna Mykich, and Igor Karpov	
Improvement and Orientation of Method of Data Arrays Sorting by Confluence on Architecture of Graphic Processor Unit	276
Ivan Tsmots, Oleksandr Kuzmin, Vasyi Dubuk, and Volodymyr Antoniv	
Capabilities and Application of Bithreshold Neurons and Networks . . .	287
Vladyslav Kotsovsky and Anatoliy Batyuk	

Intersection of Fuzzy Homogeneous Classes of Objects	306
Dmytro O. Terletskyi and Oleksandr I. Provotar	
Hybridisation of IT Project Management Methodologies. Complementary or Contradictory?	324
Sergey Bushuyev, Victoria Bushuieva, Denis Bushuiev, and Nanaliya Bushuyeva	
Big Data	
Developing Methods for Building Intelligent Systems of Information Resources Processing Using an Ontological Approach	345
Vasyl Lytvyn, Victoria Vysotska, Myroslava Bublyk, Piotr Grudowski, Yurii Matseliukh, and Roman Nanivskyi	
Adaptive Learning of Probabilistic Neural Network in Situation of Overlapping Classes in Classification Task	371
Yevgeniy Bodyanskiy, Anastasiia Deineko, Iryna Pliss, and Olha Chala	
Formulating a Geolocation Bias Correction for DMSP Nighttime Lights of Global Cities	383
Vitalii Kinakh, Tomohiro Oda, and Rostyslav Bun	
Inductive Modeling	
Liver Pathological States Identification with Self-organization Models Based on Ultrasound Images Texture Features	401
Ievgen Nastenکو, Vitaliy Maksymenko, Alexander Galkin, Vladimir Pavlov, Olena Nosovets, Irina Dykan, Boris Tarasiuk, Vitalii Babenko, Vitalii Umanets, Olena Petrunina, and Denys Klymenko	
Identification the Models of Atmospheric Pollution by Nitrogen Dioxide Based on the Artificial Bee Colony Algorithm with Modified Operators for Determining of Profitable Food Sources	419
Mykola Dyvak, Natalia Porplytsya, Libor Dostálek, Iryna Oliynyk, and Sergiy Nadvynychnyyò	
Modeling of Photovoltaic Installation Performance Taking into Account Seasonal Phenomena of Different Climate Zones	433
Natalia Porplytsya, Mykola Dyvak, Janusz Zarębski, Krzysztof Górecki, and Yurii Maslyiak	
Study of Asymptotic Behavior of a Broad Class of Criteria for Data-Driven Best Model Selection	447
Volodymyr Stepashko	
GMDH-Based Discovering Dynamic Regularities of the Ukraine Covid-19 Pandemic Process	456
Olha Moroz and Volodymyr Stepashko	

Mathematical Modeling

Mathematical Models of Nonlinear Transverse Oscillations of Elastic Movable 1D Bodies	473
Petro Pukach, Andrii Slipchuk, Halyna Beregova, and Yulia Pukach	
Methods for Calculating a Mathematical Model for Determining the Electromagnetic Field in Conducting Ferromagnetic Layer	485
Yaroslav Pelekh, Andrii Kunynets, Serhii Mentynskyi, Bohdan Fil, and Pavlo Pukach	
Modeling of Parameters of State Participation in Financing of Energy Saving Projects at Enterprises	498
Olexandr Yemelyanov, Anastasiya Symak, Liliya Lesyk, Tetyana Petrushka, Natalia Kryvinska, and Olena Vovk	
Modeling of Carbon Monoxide Oxidation on Gold Nanoparticles: Is There Oscillatory Mode?	512
Petro Kostrobij and Iryna Ryzha	
Development of the Software for the Road-Train Stable Movement Mode Research	522
Olha Sakno, Dmytro Moisia, Ievgen Medvediev, Tatiana Kolesnikova, and Andrii Rogovyi	
The Two-Point Problem as the Mathematical Model of the Oscillation Process of a Longitudinal Body	540
Zinovii Nytrebych and Oksana Malanchuk	
Study of the Anisotropy Effect in Piecewise Homogeneous Media by Boundary and Near-Boundary Element Methods	551
Liubov Zhuravchak and Nataliya Zabrodska	
Abilities of Near-Boundary and Contact Element Methods for Modeling of Non-stationary Process in Objects Under Conditions of Ideal Contact Between the Components	564
Liubov Zhuravchak	
Development and Research of Electromagnetic Installation of Industrial Wastewater Purification from Ferromagnetic Impurities as an Object of Automation	578
Andrii Safonyk, Ivan Tarhonii, Ivanna Hrytsiuk, and Olena Prysiazhniuk	
Mathematical Modeling and Program Implementation of Gasdynamic Solution of Dry Gas Seals for Centrifugal Compressors	591
Lyudmyla Rozova and Gennadii Martynenko	
Computer Simulation for Quantum Tomography	603
Bohdan Yavorsky	

Methods for Estimating the Discrete Rhythmic Structure of Cyclic Random Processes Using Adaptive Interpolation	614
Serhii Lupenko, Anatoliy Lupenko, Iaroslav Lytvynenko, and Vasyl Martsenyuk	
Theoretical Bases for Reducing the Time Complexity of the Rabin Cryptosystem	628
Mykhailo Kasianchuk, Ihor Yakymenko, Mikolaj Karpinski, Ruslan Shevchuk, Volodymyr Karpynskyi, and Inna Shylinska	
Variational Formulation of Viscoelastic Deformation Problem in Capillary-Porous Materials with Fractal Structure	640
Volodymyr Shymanskyi and Yaroslav Sokolovsky	
Mathematical Method for Processing SCADA Information and Diagnostic Flows	655
Olena Syrotkina, Mykhailo Aleksiev, Borys Moroz, Iryna Udovik, Andrii Martynenko, and Viktoriia Hnatushenko	
Assessment of Unknown Phase Shift for Speckle Interferometry Using Sample Pearson Correlation Coefficient	671
Leonid Muravsky, Yuriy Kotsiuba, and Yaroslav Kulynych	
Implementation of Piecewise Linear Stretching in Algebraic Model for Logarithmic Image Processing	682
Olena Berehulyak and Roman Vorobel	
The Quadrature Components of Narrowband Periodically Non-stationary Random Signals	696
Ihor Javorskyj, Roman Yuzefovych, Pavlo Kurapov, and Oleh Lychak	
The Study of Cellular Automata Method When Used in the Problem of Capillary-Porous Material Thermal Conductivity	714
Yaroslav Sokolovsky, Oleksiy Sinkevych, Roman Voliansky, and Volodymyr Kryshtapovych	
Simulation of Shelterwood Logging in the Global Forest Model (G4M)	730
Mykola Gusti, Fulvio Di Fulvio, and Nicklas Forsell	
Method of Adaptive Multi-scale Transformation for Image Data Compression	743
Kolganova Olena, Tereshchenko Lidiia, Alla Sitko, Kravchenko Viktoriia, Viktoriia Volkogon, Zhanna V. Vasylieva-Shalamova, and Volodymyr Shutko	
Porosity Estimation and Analysis of Images of Oxide Ceramic Coatings of D16T Alloy	758
Iryna Ivashenko, Volodymyr Posuvailo, Halyna Veselivska, and Vasyl Vynar	

Estimating the Efficiency of the Energy Service Market Functioning in Ukraine	768
Vasyl Brych, Volodymyr Manzhula, Bohdan Brych, František Drdák, Dmytro Shushpanov, and Nataliya Halys	
Econometric Pricing Model for R&D Products in Transfer Agreements	779
Vasyl Kozyk, Oleksandra Mrykhina, Lidiya Lisovska, Oksana Yurynets, and Halyna Rachynska	
Event-Based Spatially Distributed Multi-Risk Analysis	798
Maryna Zharikova and Volodymyr Sherstjuk	
Information Application for Visually Impaired People with a Ukrainian Prototype Font	814
Mariia Nazarkevych, Irina Lozovytska, Alina Zakharova, and Ivanna Klyujnyk	
Modeling the Dynamics of Computer Hardware Market Distribution	823
Nataliya Melnyk, Mykola Dyvak, Petro Stakhiv, Bohdan Melnyk, Zora Rihova, and Marta Vohnoutova	
Technology of Quantitative Integral Assessment and Forecast of a Complex Economic System Performance	841
Volodymyr Stepashko, Roman Voloschuk, and Serhiy Yefimenko	
Synthesis of Physiotherapeutic Parameters of Devices for Post-medical Restoration of Spinal Zones	857
Alexander Trunov	
Aggregation, Storing, Multidimensional Representation and Processing of COVID-19 Data	875
Oleksii Duda, Nataliia Kunanets, Oleksandr Matsiuk, Volodymyr Pasichnyk, and Antonii Rzheuskyi	
Action-Entropy Approach to Modeling of ‘Infodemic-Pandemic’ System on the COVID - 19 Cases	890
Alla Bondar, Sergey Bushuyev, Victoria Bushuieva, Nataliya Bushuyeva, and Svitlana Onyshchenko	
Project Management	
Method for Estimation of R&D Products Cost Taking into Account Their Technological Readiness	907
Nataliya Chukhray, Nataliya Shakhovska, and Oleksandra Mrykhina	

The Method and Results of Agreement of Configurations of the Integrated Projects on Agro-Industrial Production	923
Anatoliy Tryhuba, Vitaliy Boyarchuk, Inna Tryhuba, Oksana Boiarchuk, Oleh Boiarchuk, and Alla Zhelyeznyak	
Development of an Information System to Minimize the Risks of Personnel Management	939
Oleh Veres, Pavlo Ilchuk, Olha Kots, Ihor Rishnyak, and Halyna Rishniak	
Creation of a Software Development Team in Scrum Projects	959
Igor Kononenko and Hlib Sushko	
Analysis of Non-sharing, Lack and Timeliness of Information in Work Teams	972
Zora Řihová	
Geometric Distortion Correction Technique of Text Images	987
Oleksandr Tymchenko, Bohdana Havrysh, Orest Khamula, Oleksandr O. Tymchenko, and Svitlana Vasiuta	
Intelligent System of Modern Production Control Based on the Methodology of Optimal Aggregation	1004
Taisa Borovska, Dmytro Hryshyn, Iryna Kolesnyk, Victor Severilov, and Tetiana Shestakevych	
Optimal Control of Modern Production Based on Virtual Reality Statistics	1018
Taisa Borovska, Dmytro Hryshyn, Iryna Kolesnyk, and Victor Severilov	
Designing the Repository of Documentary Cultural Heritage	1034
Nataliia Kunanets, Viktoriya Dobrovolska, Nataliia Filippova, Kazimi Parviz, Halyna Lypak, Oleksii Duda, Nataliia Veretennikova, Marta Karp, Chrystyna Shunevych, and Luybov Dubrovina	
Intelligent Method of Forming the HR Management Short-Term Project	1045
Hrystyna Lipyanina, Oleg Sachenko, Taras Lendyuk, Anatoliy Sachenko, and Nadiia Vasylykiv	
Software	
Hybrid Conceptual Models, Ontologies System and Onto-Oriented Information Systems for Chinese Image Medicine as a Component of Integrative Scientific Medicine	1059
Serhii Lupenko, Oleksandra Orobchuk, Ihor Kateryniuk, and Edgars Vasilevskis	

**New Technique of Recursive Mean-Separate Contrast Stretching
for Image Enhancement 1078**
Sergei Yelmanov and Yuriy Romanyshyn

**Application of Neural Networks in Intrusion Monitoring Systems
for Wireless Sensor Networks 1101**
Olexander Belej, Kostiantyn Kolesnyk, and Orest Polotai

**Modeling an IT for Decision-Making in Education of Students
with Autism 1116**
Vasyl Andrunyk, Yurii Prystai, and Tetiana Shestakevych

**Constructive-Synthesizing Modeling of Lightning Flashes
in the Dynamic Thunderstorm Front 1128**
Viktor Shynkarenko, Iryna Nikitina, and Robert Chyhir

**A Quantitative Assessment of the Incomplete Integral Contrast
for Complex Images 1146**
Sergei Yelmanov and Yuriy Romanyshyn







**Construction of a Multisensor UAV System for Early Detection
of Forest Pests 1164**
Milan Novák, Jakub Geyer, Miloš Prokýšek, Martin Hais, Stanislav Grill,
Markéta Davidková, Petr Doležal, Peter Hofmann, and Rajan Paudyal

Author Index 1183

Applied Linguistics



Experimental Investigation of Significant Keywords Search in Ukrainian Content

Oleg Bisikalo¹ , Victoria Vysotska² , Vasyl Lytvyn² ,
Oksana Brodyak² , Svitlana Vyshemyrska³ , and Yuriy Rozov³ 

¹ Vinnytsia National Technical University, Vinnytsia, Ukraine
obisikalo@gmail.com

² Lviv Polytechnic National University, Lviv, Ukraine
{Victoria.A.Vysotska, vasyi.v.lytvyn}@lpnu.ua

³ Kherson National Technical University, Kherson, Ukraine
printvvs@gmail.com, rozov.yg@gmail.com

Abstract. The article deals with a comparative experimental study of methods of searching for significant keywords of Ukrainian-language content. The approach to the automatic definition of keywords is based on Porter's stemming of words of the Ukrainian language for the Levenshtein distance, taking into account the possibility of using a thematic dictionary and the removal of blocked words. Experimental based on 100 scientific publications of technical direction compared to the author's variants obtained numerous statistical characteristics of the accuracy of search results.

Keywords: NLP · Indexing · Keywords extraction · Classification · Ukrainian language · Porter stemming · Levenshtein distance · Ukrainian · Keywords · Search · Thematic dictionary

1 Introduction

The most effective methods of attracting potential customers and visitors to information resources in the Internet space are the proper formation and application of multiple keywords in the content provided in these information resources. One of the simplest (but painstaking and time-consuming) methods is to survey regular users and to generate a host of likes, benefits, and disadvantages (so-called ratings) of content hosted on an information resource. For example, some of our advertiser sites provide users with forms that allow them to submit their contact information, obtain additional information about the advertiser's products or services, and comment. Persons who have provided information through such forms on the advertiser's site are potential clients who can contact by telephone or email, as well as manually analyzing textual content arrays to generate statistics on the preferences and preferences of potential and/or regular audience. Another popular way is to use the Google AdWords Campaign Optimization Kit [1–7] to help you get the most out of your submitted data. Therefore, thanks to Google AdWords, they place advertisements near search results to increase traffic to the

information resource and content/product sales. After clicking on the ad, visitors expect to take some action on the landing page. It is important for the visitor to have a fair idea of what to expect before clicking an ad. To do this, they optimize their AdWords campaign, bid on the keywords they need, use negative keywords (when looking at their search terms reports), use laconic but meaningful and engaging ad text, and set up conversion tracking. Tracking the landing page of bounce rate and conversion rate, as well as comparing them to different variations of ad text that drives traffic to the page, can help you determine which ad performs better. The main question remains - how for optimally, adequately and effectively define keywords. For English-language texts, this is no longer a problem - there are many publications on the subject, software developed. For Ukrainian-language texts, similar algorithms are not effective compared to English-language ones, which is why the problem of finding Ukrainian keywords is urgent.

The article deals with the scientific and practical task of developing a method of processing Ukrainian textual information for the automatic detection of meaningful keywords and content rubrics in Internet systems. The work is performed within the framework of joint scientific researches of the Department of Information Systems and Networks of the Lviv Polytechnic National University in work "Research, development and implementation of intelligent distributed information technologies and systems based on database resources, data warehouses, data spaces and knowledge in order to accelerate the formation processes of modern information society". As well as the department of automation and information-measuring technology of Vinnitsa National Technical University within spine Research Center of Applied and Computational Linguistics. The results of the research are carried out within the framework of the state budget research works on the topics "Development of methods, algorithms and software for modeling, designing and optimization of intellectual information systems based on Web technologies, WEB" and "Intelligent information technology of image analysis of text and synthesis of integrated knowledge base - language content". Scientific researchers are also conducted within the framework of the initiative topic of the ISM department at Lviv Polytechnic National University on the topic "Development of intelligent distributed systems based on ontological approach for the integration of information resources".

2 Related Works

Choosing the right keyword list for your campaign helps you show your ads to your target customers when they search for the right information or visit certain information resources [7]. Keywords should match the queries that potential customers will search for products or services [7].

First, you should always consider the opinion of the potential user; you need to put yourself in the client's place or to analyze the questionnaires that are previously exposed on the information resource. It is necessary to define the main categories of content of the information resource (to make a list of headings), as well as the phrases and terms that characterize these categories (to make a terminal dictionary, where each word belongs to a certain rubric, or as a percentage belongs to several rubrics). This

list should also include terms or phrases that could use to describe content, products, or services. For example, many of the keywords $M = \{\text{чоловіче спортивне взуття, чоловічі кросівки, чоловіче тенісне взуття, ...}\}$ are used to describe men's sports shoes, expanding this list to both general categories that customers would use and more commonly used terms using brand names and products.

Second, you need to choose general or specific keywords depending on your goal. However, if the keywords are too specific, it will not reach all the desired target audience. Keywords that are more general are used to maximize your reach. However, keywords that are too general in scope are not always effective, because they often trigger ads for searches that have little relevance to the content of the content, reducing the chances of reaching potential audiences. In addition, there is always intense competition for such keywords.

Both the more specific and the more general keywords are used to refine the accuracy of the content description, and then determine, for example, through Google Analytics, which of them produce better results. Regardless of the level of generality or specificity of the keywords, they should always be as consistent as possible with the profile of the information resource and the general content of the content stream. This is a great way to avoid duplicate keywords in your account, since for a given keyword; Google only displays one ad per advertiser. For example, to describe men's sports shoes, it is possible to expand the set of keywords $M = \{\text{чоловіче взуття для баскетболу, дитяче спортивне взуття, спортивне взуття для тенісу, ...}\}$ to match the variety of products offered. In this case, the ad or list because of the search engine will display when someone searches for a specific type of shoes or browses the relevant thematic information resource, such as basketball or tennis. You can add general keywords (such as *взуття*) to your keyword list. In this case, your ad will appear when searching for shoes of any type, as well as on fashion related information resources.

Third, you need to group similar keywords by topic. If you add all your keywords and ads to the same ad group, a customer looking for women's evening shoes may see ads for men's tennis shoes. To show ads that are more relevant to potential customers, you need to group your keywords and ads into ad groups based on products, services, or other categories. In addition, splitting your keywords into thematic groups will make it easier to use your account. For example, a shoe storeowner can create two ad groups: one for running shoes and one for evening shoes. A running shoe ad group may include keywords *взуття для бігу* and *кросівки для бігу*, and ads targeted to people searching for running shoes. The Evening Shoes ad group may include the keywords *вечірнє взуття* and *модельні черевики* and ads targeted to people who are looking for evening shoes. Then potential customers see ads for evening shoes when they search for one of the keywords in that ad group, such *модельні черевики*.

Fourth, you need to select the right number of keywords. More often, they specify 5-20 keywords for an ad group, though more are possible. However, each ad group you create must contain keywords that are directly related to the topic of that ad group. You do not need to include other keyword variations in your ad group, such as the plural form of a keyword or keywords that may be misspelled. Phrases, which are keywords of two or three words, work best. You can select up to 20,000 individual targeting options (including keywords) in your ad group and up to 5 million in your account. However, a

small group of well-targeted keywords triggers the vast majority of relevant clicks. For example, if an ad group contains a broad match keyword *тенісне взуття*, the ad will appear when someone searches for any variation of that keyword: *тенісне взуття*, *придбати тенісне взуття*, *взуття для бігу* or *тенісні кросівки*.

3 Materials and Methods

3.1 Detailed Description of the Project

In specialized areas, such as medicine, computer sciences, agriculture, or law, terminologies play an important role and allow encoding specific knowledge from these areas, thanks to the recruitment of terms and of semantic relations among them [1–5]. Such resources are essential for several applications designed for the processing of textual data: information retrieval, information extraction, machine translation [6–13]. Yet, even if such resources are available, they are often insufficient and inadequate, and it is necessary to plan an important work for their adaptation and enrichment in order to make them exploitable in the automatic applications. Besides, terminological resources are not available for several areas and languages [14–20]. Hence, it is important to provide methodologies for the automatic building of such resources.

Since 90 s, several approaches have proposed in order to help building terminological resources from specialized texts (scientific literature, technical documentation, legal texts, etc.). These approaches permit to perform two tasks: (1) extraction of candidate terms, which are syntactic groups of words with characteristics of terms, and (2) identification of semantic relations (synonymy, hyperonymy, relations specific to the area) among these terms. These approaches have been proposed and designed mainly for texts written in English, French and Japanese. Although, more recently, an effort has paid to adapt the approaches to other European (Spanish, Italian, German, Polish, etc.) and Semitic languages. Yet, other languages, like Ukrainian, are less resourced in tools and resources, and require an important effort for filling the gap when designing such tools and creating such resources.

The general objective of the project is to propose methods, tools and resources for the corpus-based terminological acquisition and automatic indexing of scientific literature written in Ukrainian. We propose to adapt some existing approaches developed in French and English to Ukrainian language, but also to explore novel methodological directions, such as cross-language transfer as well as qualitative and quantitative approaches for semantic analysis.

The project strongly relies on expertise of researchers from Ukrainian and French teams: text mining, terminological acquisition and cross-lingual transfer for the French team, statistical methods, machine learning and acquisition of semantic relations for the Ukrainian team. Both teams are used to work with textual corpora. The main steps of the proposed methodology are presented at the Fig. 1.

The tasks of the project will be performed using specific textual material: (1) already annotated texts with syntactic information, (2) raw corpora for text mining (acquisition of lexicon and of terminology), and (3) articles of the journals for which abstracts, full texts and indexing keywords are already available. These linguistic data will be used for preparing the resources, for fitting the automatic systems and for evaluating the obtained results.

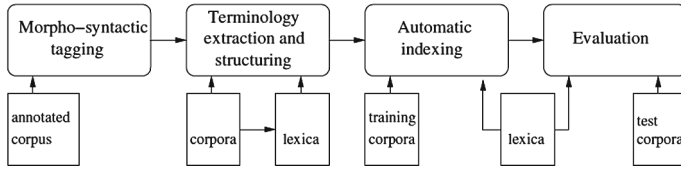


Fig. 1. General schema of the method.

The textual data, collected and available in three languages (Ukrainian, English and French) will be used. Indeed, several methods, tools and resources are already available for French and English languages, while Ukrainian language is the target language for which these methods, tools and resources will be adapted. The general method is composed of several steps essential to the realization of the project:

1. Morph-Syntactic Tagging. Morph-syntactic tagging is the first step required for preparing further steps and tasks of the project. The purpose of morph-syntactic, or Part-of-Speech (POS), tagging is to associate words from text with syntactic categories such as noun, verb, adjective, etc. together with some morphological features (gender, number, verbal tense, etc.). Indeed, this kind of information is essential for various NLP applications, such as terminology and information extraction, indexing, semantic analysis of texts. Currently, there is no usable POS-taggers in Ukrainian. For instance, the available UGtag POS-tagger does not perform the syntactic and morphological disambiguation of the tags, while the Ukrainian model for the TNT tagger remains difficult to access and to use. Hence, our first objective is to create a POS-tagger and lemmatizer dedicated to the processing of Ukrainian texts. Building of the annotated corpus has already started and several dozens of manually tagged documents with syntactic and morphological information in Ukrainian are already available. These documents cover general and specialized languages (e.g., medicine, linguistics, computer sciences). This is a very time and effort-consuming task, which has been performed by a native Ukrainian speaker with linguistic training. During the INDEX project, this annotated corpus will be exploited with supervised machine learning algorithms in order to create a syntactic model of the Ukrainian language. As noticed above, syntactic tagging provides the basic and necessary information for further steps of the project.

2. Terminology Building. Terminology building is related to the extraction of terms and to the detection of semantic relations between these terms, in order to create a thesaurus designed for indexing and for information retrieval. Such a thesaurus is necessary for the automatic indexing of scientific texts. Three main tasks are planned within this step:

- **The first task** is related to the terminology extraction. The purpose is to process the texts and to detect their linguistic units which may correspond to terms. Typically, this is done based on POS-tagging with specific shallow-parsing rules in order to detect sequences such as Adjective Noun, Noun Noun or Noun Preposition Noun, which may correspond to terms (e.g., computer sciences, information retrieval, and extraction of terms, respectively). For this task, we plan (1) to exploit crosslingual transfer approach, such as we started to do it in a previous work, and (2) to adapt

an existing term extractor YaTeA already available for French and English texts, and being currently adapted to the modern Arabic language. The YaTeA term extractor is based on specific rules designed for the shallow analysis of texts, which permits to identify textual segments (such as those presented above) which can correspond to or contain terms. Then, a syntactic analysis of these segments is done, which allows keeping only those syntactic groups, which are syntactically well formed. Besides, a first filtering of term candidates through exploitation of several statistical measures is also performed at this step. Given the application and use of YaTeA in such a multilingual context, we assume that its adaptation to Ukrainian language can be planned as part of the INDEX project. The transfer approach will exploit our first observations and results. Typically, (1) this work will rely on our current expertise in processing French and English textual documents with the YaTeA term extractor and (2) on the alignment of parallel documents at the word level, thanks to the available tools such as Giza ++ or Fast Align;

- **The second task** is related to the terminology structuring and to the detection of semantic relations between terms, such as hyperonymy or synonymy. The two involved teams have an important expertise in this research area and can efficiently contribute to this task. Moreover, qualitative and quantitative methods from the state of the art will be used for extracting the semantic relations. We will also test cross-lingual transfer approaches on parallel and comparable corpora. For instance, semantic relations acquired on French, English corpora can be projected, *and* transferred on Ukrainian terms extracted in the previous task. The result of this task will be a specific and automatically built thesaurus in Ukrainian. The terms will be structured with at least hierarchical relations, in order to permit a better knowledge representation and classification of articles. The thesaurus will be three-lingual with the corresponding terms in English and French. The content of the thesaurus will cover the computer sciences area (mathematics, informatics, etc.). These terms will be acquired from corpora and from already available resources. These terms will be enriched with synonyms thanks to the UKRWordNet resource;
- **The third task** is related to the extraction of definitions for terms. The purpose of this task is to *provide* definitions to some ambiguous terms extracted previously.

3. Automatic Indexing. The purpose of the automatic indexing of articles is to detect important keywords or key-phrases, which represent the main content of these articles. Editors or information science experts usually do this task manually. It is a time and effort-consuming task. We would like to propose original automatic methods for the detection of keywords and key-phrases, which is a very challenging topic. The development of the automatic indexing will be exploiting the available data from the journal for which abstracts, full texts and indexing keywords are already available. One subset of the completely available data, called training corpus, will be used for creating and fitting the automatic system. Hence, methods issued from supervised machine learning and information retrieval, as well as rule-based approaches can be exploited. Besides, methods for rating and filtering of the extracted keywords and key-phrases will also be used in order to select those keywords and key-phrases, which are the most suitable for indexing a given article. The automatic indexing system for scientific literature in Ukrainian will be the result of the project. The indexer will permit to create an automatic index with

main keywords of articles in Ukrainian. Its multilingual nature will also allow searching information in English and French as well. The index will be following the structure of the thesaurus and will have a hierarchical structure as well. Thanks to the keywords extracted, the articles will be associated with a given branch of the thesaurus and classified. Again, thanks to the multilingualism of the thesaurus, the indexing result will clearly state to the international community. This indexer and classificatory will allow to automatically analyze and classify scientific work of Ukrainian scientists. Hence, the indexing of their work will be available for the international audience and will permit to open new directions of the research and encourage international collaborations.

4. Evaluation. The evaluation will mainly address the main step of the project dedicated to the automatic indexing. For this, another independent test set from the journal data (abstracts, full texts and keywords) will be exploited. Classical evaluation measures will be applied (precision, recall, F-measure, MAP, etc.) in order to evaluate the performance of the indexing system.

3.2 Use the Keyword Planner, Negative Keywords, and Search Terms Report

1. Use the Keyword Planner to find new keyword ideas. With Keyword Planner, you can find new keyword ideas and get traffic estimates, which will greatly help you create a search campaign. This tool also lets you know how effective your keyword list can be, and see the average number of times people searched for those terms. Based on this data, they decide which keywords can help increase traffic to the inventory and raise awareness of the product being advertised, and more. For example, if you enter the phrase *бігове взуття* in the Keyword Planner, you might be offered additional keywords, such *бігове взуття зі знижкою* or *бігове взуття з керуванням рухом*. Statistics for each keyword variant, such as a keyword's competitiveness or average number of searches for that term, are provided for each keyword variant. These statistics help you decide which keywords to add to your list.
2. Negative keywords raise the CTR. Occasionally; you may want to prevent your ad from showing on terms that do not match your products or services. In this case, they add negative keywords to reduce costs and trigger ads only on the search terms you want. For example, for an informational resource that only sells *чоловіче взуття для бігу*, you could add the terms *жінки* and *дівчата* as negative keywords to prevent your ad from appearing when people search for the right shoes.
3. The Search terms report improves your keyword list and provides information about what people searched for when they saw and clicked on your ad. This data removes poor performing keywords and adds new ones to the list. In addition, negative keywords are defined.
4. Parser and stemming algorithms for defining keywords in plural text content [11–18, 22–30]. Parser is a parser that, usually programmatically, converts input text to a structured format. For context-dependent grammars that are native to natural languages, parsing algorithms are characterized by high complexity and poor quality - especially for synthetic languages, such as Slavic. The parsers of modern linguistic packages allow us to come to the vocabulary of not only individual words, but also the corresponding word forms, lexemes and lemmas (almost the roots of the word).

Stemming is the process of reducing a word to its base by discarding auxiliary parts, such as an ending or suffix [11–18, 31–40]. Stemming results are similar to determining the root of a word, but its algorithms are based on other principles [41–48]. Therefore, the word after processing the algorithm of stemming is often different from the morphological root of the word. Stemming is used in linguistic morphology, content analysis, and content monitoring. Search engines use stemming to combine words that match the form after the termization, so-called technical synonyms. This process is a merger. During stemming, words are fast converted to quick. In addition, the word running, running, running to the root of the word running. For the first time, the Stemming algorithm is published in [20] - it was an advanced work that had a great impact on further research in this area. Later, the Stemming algorithm is written by Martin Porter and is published in the July 1980 issue of Program. This algorithm has become widespread and has become the de facto standard stemming algorithm for English. There are quite a few implementations of the Porter algorithm that are freely distributed in software, but some of them have some disadvantages [49–53]. As a result, not all stemming algorithms produce the result they expect. To reduce such errors, Martin Porter created the official free implementation of the algorithm in 2000. In addition, over the next few years, he devoted himself to building Snowball, a special environment for writing stemming algorithms that is designed to improve English language spamming and writing more languages [54–60].

3.3 Use Keyword Match Types, Generate Multiple Keystrokes Manually, or with the Help of Stemming Algorithms

1. Use keyword-matching types to better target your ads. For example, exact match ads will only show when users search for a specific keyword or a close keyword variant (spelling or plural). Keywords are case insensitive, so it does not matter if they are case-sensitive. For example, you do not need to specify both *взуття для бігу* and *Взуття для бігу* as keywords, as one option is *взуття для бігу*. For app install campaigns, the AdWords system can expand the relevance of some keywords to the specifics of the app, including:

- **Exact and phrase match keywords** - Make minor changes (such as deleting or adding a word) to the search terms to improve the relevance of your target keywords.
- **Broad-matched keywords** - Use app category information to refine targeting and improve reach.

For example, to display only for people interested in purchasing men's running shoes, you could add the terms *чоловіче взуття для бігу* and *чоловіче бігове взуття* as exact match keywords. This way, your ad will appear when users search for exactly those terms or close variants, such as *чоловіче взуття для бігу*. It will not appear when searched for terms such as *найкраще чоловіче бігове взуття*, as this phrase contains the term best, which is not in the exact, match keyword and which is not a close variant of it.

2. Choosing keywords related to applications or information resources that your customers view. Through the Internet, your keyword list shows your ads in relevant apps or information resources that potential customers visit. Therefore, it is important to choose keywords that are related to each other as well as content that potential customers are viewing. For example, the AdWords system could expand the range of keywords to show ads for more relevant search terms. Ads are placed on relevant keyword-based information resources. Therefore, only broad match is established for all keywords. To improve your keyword performance, some of them can be excluded from ad groups. For example, for a list of keywords with shoe-related terms, shoe information resources will be targeted by the keywords in that list. You can also exclude *лижі* and *сноуборд* to prevent your ads from appearing on winter sports sites.
3. Automatically generate multiple keywords using the Stemming algorithm. There are several variants of stemming algorithms that differ in accuracy and performance: table search, flexion and suffix cutoffs, lemmatization, stochastic algorithms, hybrid approach, prefix cutoff, matching search, Ukrainian stemming [11–18, 22–30]. In the Table 1, a comparative analysis of the features, advantages and disadvantages of known stemming algorithms is performed.

Table 1. Basic stemming algorithms

Name	Feature	Example	Advantages	Disadvantages
Search by table	The table summarizes everything possible options words and their forms after the stemming.	Stemming = {інформац} → Word = {інформаційний, інформаційна, інформаційне, інформаційним, інформаційними, інформаційних, інформаційні, інформаційній, інформаційнім, інформаційного, безпритульної, інформаційному, інформаційною, інформаційну}	The simplicity, speed and convenience of handling exceptions to language rules. For languages with simple morphology (English), the tables are small.	The search table should contain all word forms. The algorithm does not work with new words. Large table sizes for languages with complex morphology (agglutinative, Slavic).
Cut off flexions and suffixes	They are based on the rules for reducing the word Rules = {Ending (ційна) → Cut (їйна); Ending (ційне) → Cut (їйне); Ending (ційний) → Cut (їйний); Ending (ційним) → Cut (їйним); Ending (ційне) → Cut (їйне)}.	Word = {інформаційний} → Stemming = {інформац}; Word = {цивілізаційний} → Stemming = {цивілізац}; Word = {приватизаційний} → Stemming = {приватизац}; Word = {культуризаційний} → Stemming = {культуризац}; Word = {національне} → Stemming = {націонал}.	The algorithm is quite compact and productive, since the number of rules is much smaller than a table with all word forms.	False conclusions and distorted forms of stemming are present (<i>пальне</i> will become <i>пал</i> instead of <i>пальні</i>). Due to the peculiarities of language, the set of rules is complicated. There are exceptions when the base words have a variable form (<i>бігом</i> and <i>біжу</i> should have <i>біз</i> , but simple clipping is not possible). This complicates the rules and adversely affects performance.
Lematization	Step 1. Defining parts of speech in a sentence (POS tagging). Step 2. The word rules apply to a part of the language..	The word <i>пальне</i> (noun) and <i>вітальне</i> (adjective) go through different chains of rules: Rules = {Ending (льне) → Cut (е); Ending (льне) → Cut (ьне)}.	The algorithms are of high quality and have a minimal error rate.	The algorithms depend on the correct recognition of parts of the language.

(continued)

Table 1. (continued)

Name	Feature	Example	Advantages	Disadvantages
Stochastic algorithms	They are based on the probability of determining the basis of a word using a knowledge base. Lemmatization has stochastic properties when it defines a part of a language without taking into account the context in which the word was used in the sentence. Preference is given to the most likely part of the language for the word.	Word = {особистість} → Stemming = {особист} → End = {ість}; Word = {сповідаю} → Stemming = {сповідаю} → End = {у}; Word = {дивимся} → Stemming = {диви} → End = {ими}, де End – the result of studying, so Word(князи) → {End(іть) = FALSE, End(у) = TRUE, End(ими) = FALSE} → Cut (у) або Word(чуйними) → {End(іть) = FALSE, End(у) = TRUE, End(ими) = TRUE} → Cut (у) OR Cut (ими).	There is only one logical rule by which we cut off the last letters from a word. Algorithms have the ability to learn and the better and larger the training base, the better the result of their work. The knowledge base for these algorithms is a set of logical rules and a lookup table.	After processing the word, several variants of the word base may emerge, from which the algorithm will select the most likely variant (prefer the stemming that shortens the word most or least). And as a result, the likelihood of stemming errors increases.
Hybrid approach	Use a combination of the above algorithms.	For example, an algorithm may use the end and suffix method, but first search the table.	The table does not contain all word forms, but exceptions to rules that are incorrectly processed by the clipping algorithm.	The probability of stemming errors increases when the rules are incorrectly described and the termination table is formed
Cut off prefixes	Trimming from words of suffixes and endings, as well as prefixes.	Word = {проголошую, наголошувати, виголошував} → Stemming = {голошу}.	В [21] детально обґрунтовано важливість такого стемінгу для деяких мов.	The probability of the formation of opposite words, i.e. Word = {незалежний} → Stemming = {залежн}.
Matching	Only use the knowledge base with the basics of the words after the stemming.	KnowledgeBase = {чорн, чорняв} → Word = {чорнява} → Count = {4, 6} → Stemming = {чорнява}. The algorithm will choose a longer variant.	Through a system of rules (the length of the match of the word and its basis) search for the most appropriate form of knowledge base	The probability of stemming errors increases when the rules are incorrectly described and the termination table is formed
Stemming in different languages	Orientation to competitive language.	If English stemming is a simple task, then Arabic or Hebrew stemming is a much more complicated task.	The first academic stemming works were devoted only to English, but now there are many implementations for other languages.	The complexity of writing stemming algorithms depends on the language features.
Stemming in Ukrainian	Stemming options for the Ukrainian language exist [1, 2] and are used in commercial search engines	–	Some steps in this direction have already been made in [6, 11–19], and the emergence of a non-commercial stemming algorithm for Ukrainian is a matter of time.	At present, there is no free implementation of such algorithms.

Two common mistakes are common in the stemming algorithms - *overstemming* and *understemming*. *Overstemming* is to reduce two different words to one basis. *Understemming* is about getting different bases for two unambiguous words with one common basis.

3.4 Highlighting Problems

Analyzing the dynamics of content flow and building the stages of processing information resources is important and relevant. Effective development and implementation of thematic information resources today is not possible without the correct definition of many keywords. In addition, they should use in a text array of content, not for some sort of distribution, but for search engines not to index and classify it as spam. Developing a method of processing Ukrainian-language text information to automatically identify meaningful keywords and rubric content is one of the strategic directions for the development of domestic e-business. Providing the ability to automate the processing of information resources to identify meaningful keywords and categorize content contributes to increasing the sales of content, products or services to a regular user, actively attracting potential users and expanding the boundaries of the target audience. In particular, these principles and technologies in e-commerce are actively used in the creation of systems of on-line/off-line sales and analysis/exchange/storage of content, e-commerce, cloud storage/computing. The lack of a common standardized approach to the processing of Ukrainian textual information in order to automatically identify meaningful keywords and content rubrics, as well as design the functionality of e-commerce support systems, raises a number of problems when implementing the typical structure of such systems.

The most common way to find keywords, check their language and write them in the appropriate table is as follows:

- 1) To overtake the word through dictionaries in a simple cycle, the main drawback is that the process will take a very long time and it will probably concluded that it is not correct to use for on-line systems;
- 2) Write a rule for the language if there are Cyrillic letters: *є, ї, і* for certain letter combinations;
- 3) If only Ukrainian and English, one can learn that one text is in Cyrillic and the other is in Latin;
- 4) The PHPLangautodetect class can define the language of a keyword;
- 5) Through Google API, but the service is paid for now;
- 6) Using the pear package to php: Text_LanguageDetect.

Analysis and processing of text arrays of data, content analysis, content search, SEO, definition of duplicates, content rubrics are all popular areas of research. In addition, in each of these areas, you need to use the keys and automatically determine the language of the texts being studied. For these tasks, the Levenstein distance (also the Levenstein function, the Levenstein algorithm or the editing distance) is used - in information theory and computational linguistics it is a measure of the difference between two sequences of characters (strings). It is calculated as the minimum number of insertion, deletion, and replacement operations required converting one sequence to another. For example, the function (`<? Echo did_you_mean (text);?>`) Corrects words using the Levenstein method [3–5], where `DB_MATRIX` is a table variable that stores the keywords.

Fuzzy search is a very useful feature of any search engine. At the same time, its effective implementation is much more complicated than simple search by exact match. The fuzzy search problem can formulated as follows: “By a given word, find in a text

or dictionary of size n all words that match the word (or begin with that word) with k possible differences.” For example, when querying «*Машина*», taking into account two possible errors, to find the words «*Маши́нка*», «*Махі́на*», «*Мали́на*», «*Кали́на*», and so on [4, 8–10]. It can also use in rubric, duplicate detection and digest formation [3, 5], for example, keyword search algorithm. Therefore, there are well-known algorithms for parsing, stemming, and keyword detection for English-language texts, but they do not properly process Ukrainian-language texts. Therefore, there is a need to adapt parsing and stemming algorithms to texts in Ukrainian, and the most promising, according to comparative analysis, may be to search for matching in conjunction with stochastic algorithms. A useful tool in the keyword search process is to use thematic vocabulary of word basics with parallel content categorization.

3.5 Formulation of the Purpose and Ideas of the Research

The purpose of this work is to determine the optimal method of automatic processing of a plurality of Ukrainian-language textual content for identifying meaningful keywords and automatic content categorization.

The processing of a plurality of C content to identify meaningful keywords is usually based on the principle of finding keywords by content (terms), based on Zipf’s law, and reduced to selecting words with an average frequency of occurrence. As this is the simplest and simplest way, it is suggested to use more sophisticated and appropriate experimental research. 100 scientific publications of the National University of Lviv Polytechnic Bulletin of the Information Systems and Networks series (<http://science.lp.edu.ua/sisn>), two issues 783 (<http://science.lp.edu.ua/SISN/SISN-2014>) and 805 (<http://science.lp.edu.ua/sisn/vol-cur-805-2014-2>) were selected as the experimental base for such research/

As it is not only necessary to organize the search for key words from a set of predefined author or moderator and known system, the idea of the research is to automatically identify from the textual data set of a set of such words that are potential key words (corresponding to certain conditions and requirements). Then the main criterion for the quality of text keyword definitions is the power of multiple-intersection sets of keystrokes - set by the author and determined automatically. Although a language-independent parser algorithm should use, the Stemming algorithm must tie to the Ukrainian language. Therefore, it is necessary to adapt the parser and stemming algorithms to the Ukrainian language using thematic vocabulary of word bases by:

- 1) first, with the help of the elementary parser algorithm (universal - the language can be any), a set of words from a text that have a certain frequency of occurrence and fall within certain significant limits, for example, 4–6%, is determined;
- 2) by means of parser and stemming algorithms, a subset of meaningful words is determined by blocking words that are moderated by the moderator into a dictionary of blocked ones, such as verbs, pronouns, prepositions, conjunctions, parts, etc.;
- 3) the new subset is compared with the content of the thematic dictionary to form a plurality of keywords (the dictionary is pre-composed by topics - each word is a heading indicator); the peculiarity is that the thematic dictionary is a dictionary of the basics of words for the Ukrainian language (for the English language is quite

simple dictionary the thematic words - the words do not change there, but in the Slavic languages there is an excess of information due to the abolition of words, the presence of prefixes, suffixes and word endings);

- 4) statistics for various texts (artistic, scientific, poetic, nonfiction, etc.) are accumulated further for the formation of subsets of keywords that will be used for the process of rubric of textual data sets;
- 5) the efficiency of the developed algorithm adapted for Ukrainian-language text arrays is compared with the known ones for different categories of texts.

4 Results Expected to Be Achieved

The INDEX project is intended to provide several results:

- proposal of approaches for the terminological acquisition from textual documents in Ukrainian;
- improvement of approaches for the terminological acquisition in French and English;
- a better knowledge of the possibilities for adaptation of terminological acquisition on other languages;
- design and test of cross-lingual transfer approaches on specialized texts in order to extract semantic information;
- collection and building of corpora and resources describing Ukrainian language, and development of methods and tools, all of which will be made freely available for the research;
- boosting the NLP research on Ukrainian language;
- sustainable exchanges between students and permanent researchers;
- a better scientific and societal cohesion between France and Ukraine;
- a better knowledge of the scientific research in computer sciences in Ukraine.

The results and knowledge generated during the project will be published and presented at international and national levels: in conference proceedings (ACL, COLING, TALN) and journal (TAL, Terminology, etc.). Resources and tools created during the project will be freely available for the scientific research.

5 Stage Plan of Works

The Gantt chart (Fig. 2) indicates the duration of the project tasks and common face-to-face meetings. The graph presents the timeframe for each of the stages described in detail in Sect. 4. The common work will be organized around bi-annual face-to-face meetings, skype conferences and exchanges of students. Skype conferences will be held at a monthly basis. They will permit to plan and follow the systematic work of participants. The face-to-face meetings will be open to other members of the laboratories and institutions involved in the project. Four common face-to-face meetings are planned during the project:

M1: kick-off meeting of the project in France. The objectives of the meeting are:

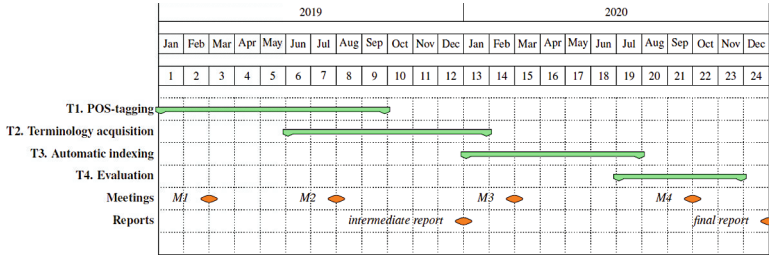


Fig. 2. Gantt chart of the project tasks.

- presentation and discussion of the objectives of the project,
- mutual presentation of the teams and people involved in the project in order to better discover scientific work and expertise of each participant,
- collection of corpora from the journals (abstracts, full texts and indexing keywords),
- planning of the work for the first year: creation of the automatic module for POS-tagging and various tasks related to the terminology building,
- seminars of Ukrainian researchers in destination to the engineer students from Paris 13 University.

M2: second meeting of the project in Ukraine. The objectives of the meeting are:

- presentation and discussion of the work done on creation of POS-tagger for Ukrainian,
- test of the POS-tagger on new documents,
- presentation of first issues on terminology acquisition,
- discussions on adaptation of the YaTeA term extractor to Ukrainian,
- planning of the work for the detection of semantic relations between terms,
- seminars of French researchers in destination to Ukrainian students.

M3: third meeting of the project in Ukraine. The objectives of the meeting are:

- test of the POS-tagger on new documents and evaluation,
- presentation of the acquired terminology (terms, semantic relations and definitions) and of the methods designed and used for its acquisition,
- planning of the work on automatic indexing (rating of terms and their filtering and selection),
- preparation of common publications,
- preparation of the intermediate report on the project realizations,
- seminars of Ukrainian researchers in destination to the engineer students.

M4: final meeting of the project in France. The objectives of the meeting are:

- presentation and discussion of work done within the frame of the project,
- preparation of the final report of the project,
- preparation of common publications,

- final workshop, possibly joined to a conference, open to other researchers for presenting the work performed by the teams involved in the project.

Final report will be prepared at the end of the project and will describe the work and realizations achieved.

6 Experiment

To achieve the goal of the study, a system (Fig. 3) is developed to select the language or languages from which the text was composed. Access to the process of finding multiple keywords based on the basics of thematic words is available at the Victana Information Resource at <http://victana.lviv.ua/index.php/kliuchovi-slova>. Developed information system for finding multiple keywords, taking into account the basics of thematic words is constructed using the following tools:

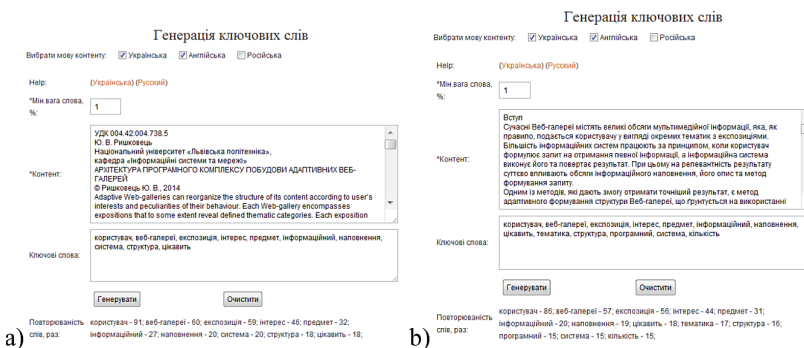


Fig. 3. Examples of article verification results

1. CMS Joomla! Version 3.4.4 to develop the e-frame of the Victana information resource,
 2. PHP version 5.3.20 to implement the algorithm of finding multiple keywords, taking into account the basics of thematic words,
 3. HTML version 4.01 to implement Web page layout,
 4. CSS to describe page styles,
 5. MySQL version 5.1.63 for data storage (dictionaries).
1. The developed information system has the following basic components.

Interface - dialog, friendly, user-friendly. Click on the *Ключові слова* (Keywords) menu item to go to the *Генерація ключових слів* (Keyword Generation) page. The page has sections (Fig. 1):

- *Вибрати мову контенту* (Select Content Language) - Select one or more languages in which the incoming text is written.

- *Help* – short instruction in Ukrainian and Russian. The instruction text opens in a separate window.
- *Мін. вага слова, %* (Minimum word weight, %) is a required field. Format - XX.XX, value - 00.01–99.99. The percentage of the keyword's weight to the total number of words in the text after which the keywords will be selected.
- *Контент* (Content) is a field in which text (article) is inserted in text format.
- *Ключові слова* (Keywords) - This field will display the keywords after clicking on the *Генерувати* (Generate) button.
- *Генерувати* (Generate) - starts the process of generating keywords when filling in the required fields.
- *Очистити* (Clear) - clears the input fields.
- *Повторюваність слів, раз* (Repeatability of words, once) - the number of repetitions of a keyword in the text.
- *Рекомендовані рубрики* (Recommended Topics) - the name speaks for itself.

2. Database (DB) - one. The tables (dictionaries) are as follows:

- Key words (keywords) - for storing and storing keywords.
- Forbidden words - To store and store forbidden words.
- headings - for the accumulation and storage of headings.
- the rules for bringing the word down - to accumulate and store rules.

All tables are edited from the administrative part. They are the same for all languages. The language attribution is defined in the table.

3. Word processing functions written in PHP.

- `blocked_words ()` function - generates a list of blocked words depending on the selected context language.
- `explode_str_on_words ()` function - clears received content from blocked words, special characters, etc.
- `count_words ()` function - calculate keyword frequency.
- `get_keywords ()` function - forming a list of keywords.
- `get_word ()` function - to write the rules for the base of the word.
- `function set_keywords ()` - write keywords to the database, if they are not there.
- `function error ()` function - error handling, sending a letter to system administrator.
- `function recommend_rubric ()` - forming a list of recommended columns.

4. Information resource in the form of Website - HTML. The page changes dynamically depending on the data entered by the user and the result of their processing, some of which are taken from the database (the keywords, headings are based on the basis).

5. CSS - used to determine the color, font, layout and other aspects of page layout.

When the resource is triggered in the * Minimum Word Weight field, % is an integer - a percentage greater than the keyword in the text. The text that you want to examine is copied to the * Content box. Clicking Generate will place a set of defined keywords in the Keywords field. Clear button required to clear the field * Content. In addition, after generating a plethora of content, Data Repeatability is displayed below the Keywords field once again - a list of found keys with numeric values (the number of words used in the text).

The analysis of statistics of functioning of the system of detection of a set of keywords from 100 scientific articles of technical direction is carried out in two stages.

1. Analyze all articles by checking the common blocked words and the thematic dictionary.
2. To analyze all articles with the check of the specified blocked words and the refined thematic dictionary as with more startup the additional unknown words (missing both in the thematic dictionary and in the set of blocked ones) are formed.

In addition, at each stage of the system verification was performed in two steps for each article: analysis of the entire article (Fig. 1a) and analysis of the article without beginning (title, authors, UDK, annotations in 2 languages, author's keywords in 2 languages, and place authors' works) and without the list of literature (Fig. 1b). This approach is used to determine the accuracy of forming multiple keywords for different modifications of the proposed method.

7 Results

Statistics are analyzed by:

- comparing multiple copyrighted keywords (defined and spelled out by the authors of these papers),
- multiple keywords defined in the first and second stages with different word weights (but more than defined in the * Min. within)

with complete and abbreviated texts (Table 2) with the arithmetic mean of the author's key phrases/words about 5 (4.77), which are formed on average from 10 (9.82) words.

The word weight is calculated as the relative frequency of occurrence of the base of the word throughout the text. In the Table 3, there are such notations as

- A is total keywords defined by the system at a given word weight,
- B is meaningful words from the list of formed, i.e. without unknown abbreviations, verbs, official words, etc.,
- C is coincidence of words with the author defined by the article,
- D is accuracy of matches found with copyrighted keywords,
- E is additional keywords defined by the system but not defined by the author of the article.

Table 2. Statistics of the studied volumes of articles

Name the volume of the article	Step 1		Step 2	
	Arithmetic mean	Total	Arithmetic mean	Total
Words	3455.8	345580	2912.47	291247
Symbols and spaces	26748.89	2674889	22659.17	2265917
Symbols	23272.09	2327209	19747.73	1974773
Rows	425.53	42553	369.65	36965
Paragraphs	164.97	16497	152.63	15263
Pages	9.56	956	8.28	828

Table 3. Statistics of the researched content of the articles

Name	Word weight	Stage 1					Stage 2				
		A	B	C	D	E	A	B	C	D	E
Step 1	≥ 5	0	0	0	0	0	0	0	0	0	0
	≥ 4	0.15	0.13	0.09	0.09	0.04	0.46	0.45	0.33	0.31	0.15
	≥ 3	0.41	0.38	0.22	0.21	0.16	1.21	1.2	0.85	0.79	0.41
	≥ 2	1.08	0.88	0.63	0.59	0.26	2.67	2.64	1.65	1.54	1.12
	≥ 1	5.46	3.92	2.51	2.08	1.74	7.43	7.03	3.27	3	4.18
Step 2	≥ 5	0.11	0.1	0.06	0.06	0.04	0.33	0.32	0.25	0.23	0.1
	≥ 4	0.19	0.17	0.12	0.12	0.05	0.73	0.72	0.45	0.42	0.31
	≥ 3	0.51	0.45	0.29	0.27	0.17	1.42	1.4	0.93	0.85	0.54
	≥ 2	1.34	1.11	0.74	0.72	0.39	3.12	3.07	1.81	1.67	1.43
	≥ 1	6.51	5.02	2.68	2.23	2.37	8.35	7.78	3.25	2.91	4.99

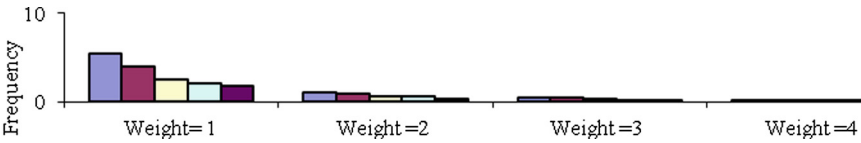


Fig. 4. Obtaining Meaningful Words When Processing Text in Stage 1, Step 1 (left to right - total words, semantic words, copyright match, match accuracy, additional words)

In Fig. 4, 5, 6, 7, 8 and 9 presents diagrams of analysis of statistics of formation by system of sets of keywords.

Figure 10a shows a diagram of the analysis of the statistics of the formation by the system of sets of all potential keywords is presented in comparison with the set defined by the authors of articles. The first column is the average number of keywords identified

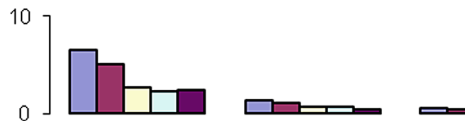


Fig. 5. Obtaining Meaningful Words When Processing Text in Stage 1, Step 2 (left to right - total words, semantic words, copyright match, match accuracy, additional words)

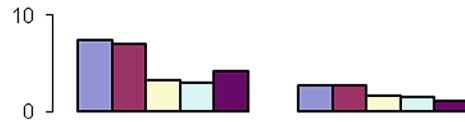


Fig. 6. Obtaining Meaningful Words When Processing Text in Stage 2, Step 1 (left to right - total words, semantic words, copyright match, match accuracy, additional words)

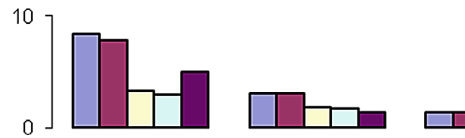


Fig. 7. Obtaining Meaningful Words When Processing Text in Stage 2, Step 2 (left to right - total words, semantic words, copyright match, match accuracy, additional words)



Fig. 8. Arithmetic mean of significant words in the text compared to the original words for stage 1 - a) step 1 and b) step 2 (left to right - copyright keywords, number of words, system-defined keywords, content words, copyright match, match accuracy, additional words)

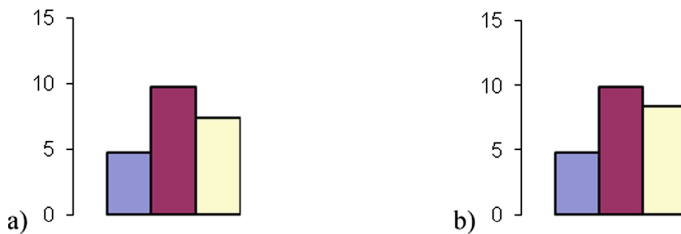


Fig. 9. Arithmetic mean of significant words in the text compared to the original words for stage 2 - a) step 1 and b) step 2 (left to right - copyright keywords, number of words, system-defined keywords, content words, copyright match, match accuracy, additional words)

by the author, and the second is the average number of words that make up those author keywords. The third column is the arithmetic average of potential keywords identified systematically in Stage 1, Step 1; the fourth is in Stage 1, step 2; the fifth is in Stage 2, step 1; sixth - in Stage 2, step 2.

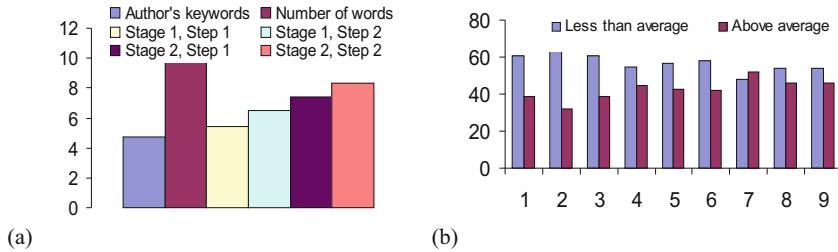


Fig. 10. Results of the review of 100 articles

Figure 10b shows the diagram of analysis of the statistics of distribution of the density of the text in the analyzed articles is presented. Number 1 is the analysis of the number of pages of articles respectively smaller and greater than the average value, 2 - paragraphs in the article, 3 - lines with text, 4 - words, 5 - characters, 6 - characters and spaces, 7 - words on the page, 8 - characters on the page, 9 - characters and spaces on the page.

8 Discussions

In Fig. 11 shows the distribution diagram of the set of sets of all potential keywords for each article compared to the set defined by the authors of the articles.

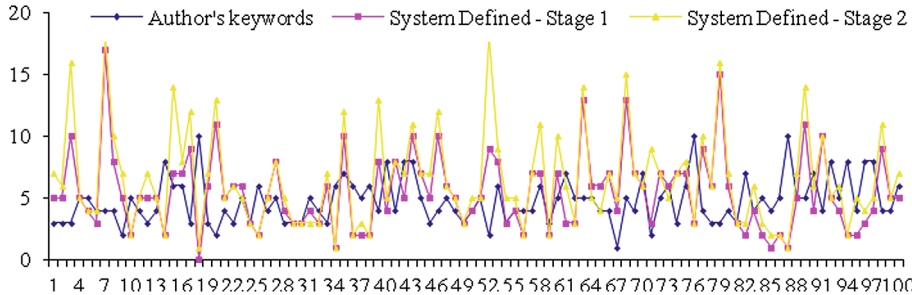


Fig. 11. Results of the review of 100 articles

The Tables 4, 5 and 6 shows the results of an analysis of the statistics of the set of sets of all potential keywords for each article compared to the set defined by the authors of the articles are presented. Symbol A is for the author's keywords, B is for the system-defined keywords in stage 1 (step 1), C is for the key words system-defined keywords in stage 1 (step 2), D is for system-defined keywords in stage 2 (step 1), E is for system-defined

Table 4. Descriptive statistics for keyword formation for the texts studied

Standard error	0.1808594	0.3103929	0.3903499	0.3012972	0.3246112
Median	4	5	6	7	8
Moda	4	5	5	7	8
Standard deviation	1.7995281	3.0883707	3.8839324	2.9978691	3.2298405
Sample dispersion	3.2383014	9.5380334	15.084931	8.9872191	10.43187
Kurtosis	0.6528151	1.7052728	0.7486433	-0.456455	-0.504378
Asymmetry	0.9479385	1.1253053	1.0657159	0.5375984	0.5170473
Interval	8	16	17	12	13
Minimum	2	1	1	2	3
Maximum	10	17	18	14	16
Sum	476	546	650	743	835
Account	99	99	99	99	99
The biggest(1)	10	17	18	14	16
The smallest(1)	2	1	1	2	3
Reliability level (95.0%)	0.3589095	0.6159647	0.7746366	0.5979144	0.6441803

Table 5. Statistics for histograms for group *A* and group *B*

<i>A</i>	Frequency	Integral %	<i>A</i>	Frequency	Integral %	<i>B</i>	Frequency	Integral %	<i>B</i>	Frequency	Integral %
1	0	0.00%	4	27	27.27%	1	2	2.02%	5	20	20.20%
2	4	4.04%	5	21	48.48%	2	10	12.12%	7	16	36.36%
3	20	24.24%	3	20	68.69%	3	12	24.24%	3	12	48.48%
4	27	51.52%	6	11	79.80%	4	4	28.28%	2	10	58.59%
5	21	72.73%	8	8	87.88%	5	20	48.48%	6	9	67.68%
6	11	83.84%	7	5	92.93%	6	9	57.58%	4	4	71.72%
7	5	88.89%	2	4	96.97%	7	16	73.74%	8	4	75.76%
8	8	96.97%	10	3	100.00%	8	4	77.78%	10	4	79.80%
9	0	96.97%	1	0	100.00%	9	2	79.80%	11	3	82.83%
10	3	100.00%	9	0	100.00%	10	4	83.84%	12	3	85.86%
11	0	100.00%	11	0	100.00%	11	3	86.87%	14	3	88.89%
12	0	100.00%	12	0	100.00%	12	3	89.90%	1	2	90.91%
13	0	100.00%	13	0	100.00%	13	2	91.92%	9	2	92.93%
14	0	100.00%	14	0	100.00%	14	3	94.95%	13	2	94.95%
15	0	100.00%	15	0	100.00%	15	1	95.96%	16	2	96.97%
16	0	100.00%	16	0	100.00%	16	2	97.98%	18	2	98.99%
17	0	100.00%	17	0	100.00%	17	0	97.98%	15	1	100.00%
18	0	100.00%	18	0	100.00%	18	2	100.00%	17	0	100.00%
More	0	100.00%	More	0	100.00%	More	0	100.00%	More	0	100.00%

keywords in stage 2 (step 2). Figure 12 and 13 the statistics of the analysis of the text of the articles in the formation of sets of keywords for the construction of appropriate histograms for groups A-E are presented.

Table 6. Statistics for histograms for group C, group D and group E

C	Frequency	Integral %	C	Frequency	Integral %	D	Frequency	Integral %	D	Frequency	Integral %
1	2	2,02%	5	20	20,20%	1	0	0.00%	7	15	15.15%
2	10	12,12%	7	16	36,36%	2	1	1.01%	6	14	29.29%
3	12	24,24%	3	12	48,48%	3	5	6.06%	5	13	42.42%
4	4	28,28%	2	10	58,59%	4	9	15.15%	10	12	54.55%
5	20	48,48%	6	9	67,68%	5	13	28.28%	8	11	65.66%
6	9	57,58%	4	4	71,72%	6	14	42.42%	4	9	74.75%
7	16	73,74%	8	4	75,76%	7	15	57.58%	12	6	80.81%
8	4	77,78%	10	4	79,80%	8	11	68.69%	3	5	85.86%
9	2	79,80%	11	3	82,83%	9	4	72.73%	14	5	90.91%
10	4	83,84%	12	3	85,86%	10	12	84.85%	9	4	94.95%
11	3	86,87%	14	3	88,89%	11	1	85.86%	13	3	97.98%
12	3	89,90%	1	2	90,91%	12	6	91.92%	2	1	98.99%
13	2	91,92%	9	2	92,93%	13	3	94.95%	11	1	100.00%
14	3	94,95%	13	2	94,95%	14	5	100.00%	1	0	100.00%
15	1	95,96%	16	2	96,97%	15	0	100.00%	15	0	100.00%
16	2	97,98%	18	2	98,99%	16	0	100.00%	16	0	100.00%
17	0	97,98%	15	1	100,00%	17	0	100.00%	17	0	100.00%
18	2	100,00%	17	0	100,00%	18	0	100.00%	18	0	100.00%
More	0	100,00%	More	0	100,00%	More	0	100.00%	More	0	100.00%

E	Frequency	Integral %	E	Frequency	Integral %
1	0	0.00%	8	14	14.14%
2	0	0.00%	5	12	26.26%
3	1	1.01%	7	11	37.37%
4	9	10.10%	10	11	48.48%
5	12	22.22%	4	9	57.58%
6	9	31.31%	6	9	66.67%
7	11	42.42%	9	9	75.76%
8	14	56.57%	11	5	80.81%
9	9	65.66%	14	5	85.86%
10	11	76.77%	12	4	89.90%
11	5	81.82%	13	4	93.94%
12	4	85.86%	15	3	96.97%
13	4	89.90%	16	2	98.99%
14	5	94.95%	3	1	100.00%
15	3	97.98%	1	0	100.00%
16	2	100.00%	2	0	100.00%
17	0	100.00%	17	0	100.00%
18	0	100.00%	18	0	100.00%
More	0	100.00%	More	0	100.00%

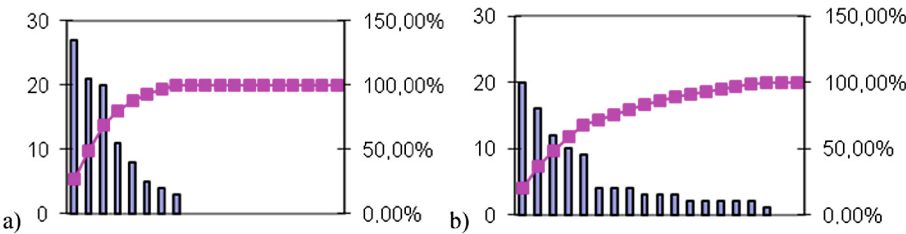


Fig. 12. Histogram for a) sample A and b) sample B

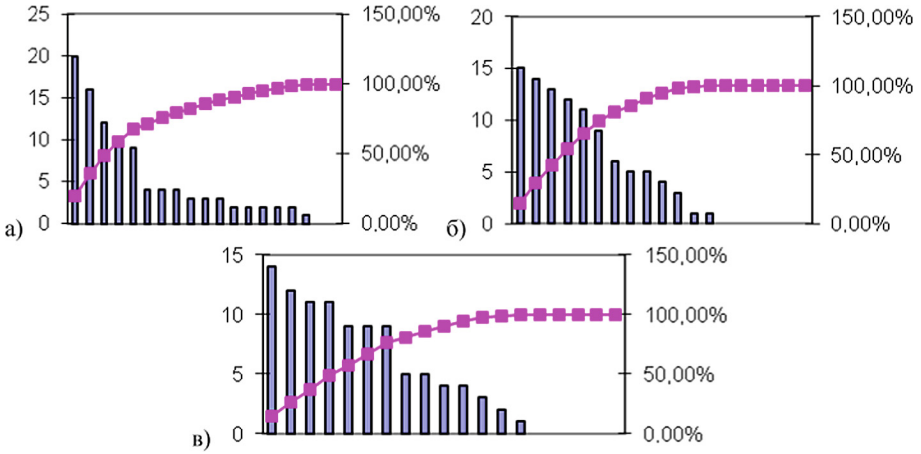


Fig. 13. Histogram for a) sample C, б) sample D and в) sample E

9 Conclusions

The objective of our project is to contribute to the evolution of the situation and to the development of novel IA-based tools and resources for Ukrainian language. More particularly, we aim at developing methods, tools and resources for performing terminology extraction and structuring starting from full text documents, and for indexing of scientific articles with suitable keywords. Such objectives are challenging not only because the project must address Ukrainian language with currently poor resources and tools, but also because the current pitfalls must be also addressed and resolved, such as terminological variation, detection of semantic relations between terms, or rating and filtering of candidates for keywords. Yet, such objectives will permit to create NLP technologies and tools (indexing, extraction of keywords, classification...) for providing direct access to Ukrainian scientific literature for researchers from other countries. Thus, our project may provide a mean for cross-lingual retrieval of scientific information. Development of technologies and tools for indexing and classification of scientific work according to international standards is an important and necessary condition for the further development of the automatic and semantic analysis of scientific literature. At a methodological level, our objective is to use cross-lingual transfer methods, which permit to transfer information extracted in texts of the source language on the texts in the target language, in order to identify there the same type of information. These approaches, which have used for some years on the problematics related to the general language, have the objective to propose solutions in several target languages in which low or no NLP resources and tools are available. For instance, this kind of approaches permitted to create methods for morph-syntactic tagging and syntactic analysis in several European low-resourced languages and for Chinese. Yet, the main difficulty with such approaches is that they require the availability or creation of parallel corpora aligned at the sentence level, which are quite rare in specialized languages. Still, these approaches provide the possibility to use existing methods and tools developed for the source language and to transpose the extracted information on the target language. In this relation, notice also that another approach exploits a machine translation system for transferring

semantic relations acquired in English on other languages. Yet another objective of the project is to make the developed methods, tools and resources freely available for scientific research. The article investigates methods of automatic detection of meaningful keywords of Ukrainian-language content based on Potter's stemming of Levenstein distances. 100 scientific publications of the National University "Lviv Polytechnic" Bulletin of the Information Systems and Networks series (<http://science.lp.edu.ua/sisn>) from two issues 783 (<http://science.lp.edu.ua/SISN/SISN-2014>) and 805 (<http://science.lp.edu.ua/sisn/vol-cur-805-2014-2>) were selected as the experimental base for this research. On the selected experimental basis, using the implemented Potter's stemming algorithm, the information resource Victana.lviv.ua obtained statistical characteristics of 4 methods:

- analyzing the entire article with a check on common blocked words and a thematic dictionary;
- analysis of the article without beginning (title, authors, udk, annotations in two languages, author's keywords in two languages, place of work of authors) and without a list of literature with a check of common blocked words and thematic dictionary;
- analysis of the entire article by checking the refined blocked words and refined thematic vocabulary (with more startup, a set of unknown words is missing, which is missing both in the thematic dictionary and in the many blocked ones);
- analysis of the article without beginning (title, authors, udk, annotations in two languages, author's keywords in two languages, place of work of the authors) and without the list of literature with verification of refined blocked words and refined thematic dictionary.

It is revealed that for technical scientific texts of the experimental base the best results are achieved by the fourth method of analysis of the article (without beginning and without the list of literature with the check of the refined blocked words and the refined thematic dictionary). This method of keyword definition is more accurate (by the vast majority of numerical metrics) and correct (the keywords found more accurately describe the subject area of the article and define the rubric of this work).

Further experimental research is needed to search for keywords in other categories of texts - artistic, non-fiction, scientific, humanities, etc.

References

1. Khomytska, I., Teslyuk, V.: Authorship and style attribution by statistical methods of style differentiation on the phonological level. In: *Advances in Intelligent Systems and Computing III*. AISC 871, pp. 105–118. Springer (2019)
2. Khomytska, I., Teslyuk, V., Holovatyy, A., Morushko, O.: Development of methods, models, and means for the author attribution of a text. *Eastern-Eur. J. Enterpr. Technol.* **3**(2–93), 41–46 (2018)
3. Cherednichenko, O., Babkova, N., Kanishcheva, O.: Complex term identification for ukrainian medical texts. In: *CEUR Workshop Proceedings*, pp. 146–154 (2018)
4. Sharonova, N., Doroshenko, A., Cherednichenko, O.: Issues of fact-based information analysis. In: *CEUR Workshop Proceedings*, vol. 2136, pp. 11–19 (2018)




5. Bobicev, V., Kanishcheva, O., Cherednichenko, O.: Sentiment analysis in the Ukrainian and Russian news. In: First Ukraine Conference on Electrical and Computer Engineering, pp. 1050–1055 (2017)
6. Vysotska, V., Burov, Y., Lytvyn, V., Demchuk, A.: Defining author's style for plagiarism detection in academic environment. In: Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP, pp. 128–133 (2018)
7. Lytvyn, V., Vysotska, V., Burov, Y., Bobyk, I., Ohirko, O.: The linguometric approach for co-authoring author's style definition. In: Intelligent Data Acquisition and Advanced Computing Systems, IDAACS-SWS, pp. 29–34 (2018)
8. Lytvyn, V., Vysotska, V., Peleshchak, I., Basyuk, T., Kovalchuk, V., Kubinska, S., Chyrun, L., Rusyn, B., Pohreliuk, L., Salo, T.: Identifying textual content based on thematic analysis of similar texts in big data. In: International Scientific and Technical Conference on Computer Science and Information Nechnologies (CSIT), pp. 84–91 (2019)
9. Babichev, S.: An evaluation of the information technology of gene expression profiles processing stability for different levels of noise components. *Data*, **3**(4), 48 (2018)
10. Babichev, S., Durnyak, B., Pikh, I., Senkivskyy, V.: An evaluation of the objective clustering inductive technology effectiveness implemented using density-based and agglomerative hierarchical clustering algorithms. In: *Advances in Intelligent Systems and Computing*, vol. 1020, pp. 532–553 (2020)
11. Senyk, M.: The Porter Stemming Algorithm for Ukrainian, <http://www.senyk.poltava.ua>, last accessed 2020/03/21
12. Vysotska, V., Lytvyn, V., Kovalchuk, V., Kubinska, S., Dilai, M., Rusyn, B., Pohreliuk, L., Chyrun, L., Chyrun, S., Brodyak, O.: Method of similar textual content selection based on thematic information retrieval. In: International Scientific and Technical Conference on Computer Science and Information Nechnologies (CSIT), pp. 1–6 (2019)
13. Vysotska, V., Fernandes, V.B., Lytvyn, V., Emmerich, M., Hrendus, M.: Method for determining linguometric coefficient dynamics of Ukrainian text content authorship. In: *Advances in Intelligent Systems and Computing*, vol. 871, pp. 132–151 (2019)
14. Lytvyn, V., Vysotska, V., Pukach, P., Nytrebych, Z., Demkiv, I., Senyk, A., Malanchuk, O., Sachenko, S., Kovalchuk, R., Huzyk, N.: Analysis of the developed quantitative method for automatic attribution of scientific and technical text content written in Ukrainian. *Eastern-Eur. J. Enterp. Technol.* **6**(2–96), 19–31 (2018)
15. Vysotska, V., Lytvyn, V., Hrendus, M., Kubinska, S., Brodyak, O.: Method of textual information authorship analysis based on stylometry. In: 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, pp. 9–16 (2018)
16. Vysotska, V., Kanishcheva, O., Hlavcheva, Y.: Authorship identification of the scientific text in Ukrainian with using the lingvometry methods. In: *Computer Sciences and Information Technologies*, CSIT, pp. 34–38 (2018)
17. Kulchytskyi, I.: Statistical analysis of the short stories by roman ivanychuk. In: *CEUR Workshop Proceedings*, vol. 2362, pp. 312–321 (2019)
18. Shandruk, U.: Quantitative characteristics of key words in texts of scientific genre (on the Material of the Ukrainian scientific journal). In: *CEUR Workshop Proceedings*, vol. 2362, pp. 163–172 (2019)
19. Hardcoded stemmer for Ukrainian. <https://github.com/vgrichina/ukrainian-stemmer>. Accessed 21 Mar 2020
20. Lovins, J.B.: Development of a stemming algorithm. *Mech. Transl. Comput. Linguist.* **11**, 22–31 (1968)
21. Jongejan, B., Dalianis, H.: Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. <http://www.aclweb.org/anthology/P/P09/P09-1017.pdf>. Accessed 21 Mar 2020

22. Moseichuk, V.: Porter stemming algorithm for Ukrainian languages. http://www.marazm.org.ua/document/stemer_ua/. Accessed 21 Mar 2020
23. Perestoronin, P.: The Porter Stemming Algorithm for Russian. <http://blog.eigene.in/post/49598738049/snowball>. Accessed 21 Mar 2020
24. Porter stemmer. <https://github.com/allaud/porter-stemmer>. Accessed 21 Mar 2020
25. Porter, M.F.: An algorithm for suffix stripping. http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html. Accessed 21 Mar 2020
26. Russian stemming algorithm. <http://snowball.tartarus.org>. Accessed 21 Mar 2020
27. The Porter Stemming Algorithm. <http://tartarus.org/~martin/PorterStemmer/>. Accessed 21 Mar 2020
28. Porter Stemming Algorithm. <http://snowball.tartarus.org/algorithms/porter/stemmer.html>. Accessed 21 Mar 2020
29. English stemming algorithm. <http://snowball.tartarus.org/algorithms/english/stemmer.html>. Accessed 21 Mar 2020
30. Willett, P.: The Porter stemming algorithm: then and now. <http://eprints.whiterose.ac.uk/1434/>. Accessed 21 Mar 2020
31. Khribi, M.K., Jemni, M., Nasraoui, O.: Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. In: International Conference on Advanced Learning Technologies, pp. 241–245 (2008)
32. Mobasher, B.: Data mining for web personalization. In: The Adaptive Web, pp. 90–135. Springer (2007)
33. Ferretti, S., Mirri, S., Prandi, C., Salomoni, P.: Automatic web content personalization through reinforcement learning. *J. Syst. Softw.* **121**, 157–169 (2016)
34. Lavie, T., Sela, M., Oppenheim, I., Inbar, O., Meyer, J.: User attitudes towards news content personalization. *Int. J. Hum.-Comput. Stud.* **68**(8), 483–495 (2010)
35. Fredrikson, M., Livshits, B. Repriv: Re-imagining content personalization and in-browser privacy. In: Symposium on Security and Privacy, pp. 131–146 (2011)
36. Chang, C.C., Chen, P.L., Chiu, F.R., Chen, Y.K.: Application of neural networks and Kano's method to content recommendation in web personalization. *Expert Syst. Appl.* **36**(3), 5310–5316 (2009)
37. Oliinyk, V.-A., Vysotska, V., Burov, Y., Mykich, K., Basto-Fernandes, V.: Propaganda detection in text data based on NLP and machine learning. In: CEUR Workshop Proceedings, vol. 2631, pp. 132–144 (2020)
38. Lynnyk, R., Vysotska, V., Matseliukh, Y., Burov, Y., Demkiv, L., Zaverbnyj, A., Sachenko, A., Shylinska, I., Yevseyeva, I., Bihun, O.: DDOS attacks analysis based on machine learning in challenges of global changes. In: CEUR Workshop Proceedings, vol. 2631, pp. 159–171 (2020)
39. Anisimova, O., Vasylenko, V., Fedushko, S.: Social networks as a tool for a higher education institution image creation. In: CEUR Workshop Proceedings, vol. 2392, pp. 54–65 (2019)
40. Antonyuk, N., Medykovskyy, M., Chyrun, L., Dverii, M., Oborska, O., Krylyshyn, M., Vysotsky, A., Tsiura, N., Naum, O.: Online tourism system development for searching and planning trips with user's requirements. In: Advances in Intelligent Systems and Computing IV, Springer Nature Switzerland AG 2020, vol. 1080, pp. 831–863 (2020)
41. Rzhеuskyi, A., Kutjuk, O., Voloshyn, O., Kowalska-Styczen, A., Voloshyn, V., Chyrun, L., Chyrun, S., Peleshko, D., Rak, T.: The intellectual system development of distant competencies analyzing for IT recruitment. In: Advances in Intelligent Systems and Computing IV, vol. 1080, pp. 696–720. Springer, Cham (2020)
42. Antonyuk, N., Chyrun, L., Andrunyk, V., Vasevych, A., Chyrun, S., Gozhyj, A., Kalinina, I., Borzov, Y.: Medical news aggregation and ranking of taking into account the user needs. In: CEUR Workshop Proceedings, vol. 2362, pp. 369–382 (2019)

43. Chyrun, L., Chyrun, L., Kis, Y., Rybak, L.: Automated information system for connection to the access point with encryption WPA2 enterprise. In: *Lecture Notes in Computational Intelligence and Decision Making*, vol. 1020, pp. 389–404 (2020)
44. Kis, Y., Chyrun, L., Tsymbaliak, T., Chyrun, L.: Development of system for managers relationship management with customers. In: *Lecture Notes in Computational Intelligence and Decision Making*, vol. 1020, pp. 405–421 (2020)
45. Chyrun, L., Kowalska-Styczen, A., Burov, Y., Berko, A., Vasevych, A., Pelekh, I., Ryshkovets, Y.: Heterogeneous data with agreed content aggregation system development. In: *CEUR Workshop Proceedings*, vol. 2386, pp. 35–54 (2019)
46. Chyrun, L., Burov, Y., Rusyn, B., Pohreliuk, L., Oleshek, O., Gozhyj, A., Bobyk, I.: Web resource changes monitoring system development. In: *CEUR Workshop Proceedings*, vol. 2386, pp. 255–273 (2019)
47. Gozhyj, A., Chyrun, L., Kowalska-Styczen, A., Lozynska, O.: Uniform method of operative content management in web systems. In: *CEUR Workshop Proceedings*, vol. 2136, pp. 62–77 (2018)
48. Chyrun, L., Gozhyj, A., Yevseyeva, I., Dosyn, D., Tyhonov, V., Zakharchuk, M.: Web content monitoring system development. In: *CEUR Workshop Proceedings*, vol. 2362, pp. 126–142 (2019)
49. Bisikalo, O., Kontsevoi, A.: System for definition of indicator characteristics of social networks participants profiles. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 77–88 (2020)
50. Kulchytskyi, I.: Quantitative parameters of some novellas by roman ivanychuk. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 89–105 (2020)
51. Levchenko, O., Tyshchenko, O., Dilai, M.: Associative verbal network of the conceptual domain БІЛІА (MISERY) in Ukrainian. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 106–120. (2020)
52. Vasyliuk, V., Shyika, Y., Shestakevych, T.: Information system of psycholinguistic text analysis. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 178–188 (2020)
53. Khomytska, I., Teslyuk, V.: The multifactor method applied for authorship attribution on the phonological level. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 189–198 (2020)
54. Albota, S.: Resolving conflict situations in reddit community driven discussion platform. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 215–226 (2020)
55. Stasiuk, L.: Computer sampling and quantitative analysis in exploring secondary functions of questions in speech genres of intimate communication. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 227–238 (2020)
56. Artemenko, O., Pasichnyk, V., Kunanets, N., Shunevych, K.: Using sentiment text analysis of user reviews in social media for e-tourism mobile recommender systems. In: *CEUR Workshop Proceedings*, vol. 2604, 259–271 (2020)
57. Bekesh, R., Chyrun, L., Kravets, P., Demchuk, A., Matseliukh, Y., Batiuk, T., Peleshchak, I., Bigun, R., Maiba, I.: Structural modeling of technical text analysis and synthesis processes. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 562–589 (2020)
58. Chyrun, L.: Model of adaptive language synthesis based on cosine conversion furries with the use of continuous fractions. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 600–611 (2020)
59. Husak, V., Lozynska, O., Karpov, I., Peleshchak, I., Chyrun, S., Vysotskyi, A.: Information system for recommendation list formation of clothes style image selection according to user's needs based on NLP and Chatbots. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 788–818 (2020)
60. Makara, S., Chyrun, L., Burov, Y., Rybchak, Z., Peleshchak, I., Peleshchak, R., Holoshchuk, R., Kubinska, S., Dmytriv, A.: An intelligent system for generating end-user symptom recommendations based on machine learning technology. In: *CEUR Workshop Proceedings*, vol. 2604, pp. 844–883 (2020)



Linguistic Analysis of Results of Variable Courses Selection by HEI's Students

Pavlo Zhezhnych  and Anna Shilinh  

Department of Social Communication and Information Activities, Lviv Polytechnic National University, 12 S. Bandery str, Lviv 79000, Ukraine
{pavlo.i.zhezhnych, Anna.Y.Shilinh}@lpnu.ua

Abstract. The aim of this article is linguistic analysis of results of variable courses selection by HEI's students. Selection of variable courses allows students to participate in the formation of ways of their individual professional development. The article presents a linguistic analysis of selected courses by students of the Lviv Polytechnic National University. It is established that students do not always pay attention to the information content of various courses, and make selection only based on their title. In the linguistic analysis process was found that most students selected courses that contain, popular for this specialty, keywords (eg, "startup", "hacking", etc.) or words that characterize course in terms of organizing their own business (eg, "business analysis", "business management", etc.). The article analyzes popular search queries and popular variable courses, taking into account their subject markers. That is shows the existence of trends dependence to selection of these courses on interest over time of search queries that contain these subject markers. The paper also establishes the accordance of popular variable courses to the main trends of information retrieval on the Internet. That is why the linguistic analysis of the results of variable courses selection will allow the representatives of higher education institutions to focus on the needs of consumers of educational services and the growing popularity of certain courses of the student's selection. This will allow HEI to effectively plan the educational process.

Keywords: Variable course · Search query · Higher education institution · Subject marker · Student · Related Query

1 Introduction

Higher education institutions (HEI) are the center of formation of modern youth and the initial stage of forming the experience of the young generation to live in market conditions, when a successful choice forms the basis for their implementation in the professional sphere. HEI educational programs take into account the individual characteristics of the consumers' development of educational services and provide them with a choice of in-depth unified professional development. In particular, such an opportunity is provided by a variable component of educational training of specialists in the chosen field. Variable courses of professional and practical training enable in-depth training

on specialties and specializations. These courses determine an essence of future activities, to promote the academic mobility of the student and his/her personal interests, to allow specializations in the basic specialty in order to form the competence of the applicant in accordance with the requirements of the labor market. Thus, studying in same course in same specialty, students have the opportunity to develop individually and adjust the acquisition of knowledge, skills and abilities in accordance with their interests and desires for implementation in the professional sphere.

However, the decision-making process by students is not always thorough in determining the necessary knowledge, skills and abilities that will agree in the future. This is mainly due to the lack of experience in the professional field, as well as age and psychological characteristics of consumers of educational services.

That is why, the linguistic analysis of results of variable courses selection by HEI's seekers will allow the representatives of higher education institutions to focus on the needs of consumers of educational services and the growing popularity of certain courses of the student's selection. This will allow HEI to effectively plan the educational process taking into account the interests of consumers of educational services in today's market conditions.

2 Related Works

Selection of a system of indicators to assess the quality of educational services provided by the institution and the mechanisms of formation of social conventions in higher education are considered in studies [1, 2]. This is a necessary condition for ensuring its quality.

Foreign experience of effective planning of the educational process are presented in [3–5].

Features of trends in potential consumers' selection of educational services are the subject of study in [6, 7]. In particular, the profession selection as a choice of life is described in [8]. In work [9], it was found, that the general motivation for professional development is based primarily on the ability to get a high-paying job. Instead, long-term goals, such as personal development and promotion, had the lowest impact on youth.

International experience in selection professional qualities for students of higher education institutions based on gender differences of students is described in [10–12]. Analysis of the level of career maturity among young people of different ages is the subject of the study [7].

The account of motivational intentions of consumers of educational services and formation of the offer of educational services of HEI based on the linguistic analysis is described in works [13, 14].

The study results of patterns, characteristics and dependencies of automatic word processing, lighting, formalization of data and information flows in the processes of content transformation are described in the study [15], and methods of linguistic analysis to automate all content processing operations are presented in [16–18].

In particular, the approaches to the classification of text documents using the ontological approach and the method of categorization of text documents on the basis of metrics, which uses the specifics of the ontology of the rubric, are constructed in [19, 20].

Solving the problem of automatic removal of key phrases from the text body related to a specific domain, so that the texts associated with common keyword phrases form a well-connected graph are presented in [21, 22]. The authors have developed a new method that uses a combination of a well-known keyword extraction algorithm (eg TextRank, Topical PageRank, KEA, Maui) using a thesaurus-based procedure that improves the connection of text graphics via the keyboard while improving the quality of extracted keywords from the point view of accuracy and reminders.

In general, all studies either establish requirements for the planning of the educational process for the future professional development of specialists by the HEI, or reveal the linguistic patterns and motivational intentions of consumers of educational services without taking into account the educational process in which they are involved. None of the studies uses linguistic analysis of variable courses selection by higher education students to take into account popular trends in professional development in the market of educational services and effective planning of educational services by higher education institutions, which forms the relevance of this study.

3 Linguistic Analysis of Results of Variable Courses Selection by HEI's Students

Today, higher education institutions are the main entities that form and provide educational services.

Requirements for courses selection, their total amount of the total number of ECTS credits, which are provided for this level of higher education, are provided by the educational program and curriculum, which are determined by the main state regulations.

In the general case, the offer of educational services has the form [15]:

$$EduProposition_i = \langle EntryYear, Spec_i, EduLevel, EduForm, Department, Course, \rangle, \\ LicQuantity_i, EduProgram_i \quad (1)$$

where *EntryYear* is the year of entry, *Spec_i* is the i-th specialty in the list of training of HEI, *EduLevel* is an educational qualification level, *EduForm* is the form of education, *Department* is the subdivision of HEI, that provides training, *Course* is the list of subjects to study, *LicQuantity_i* is the license volume of the i-th specialty, *EduProgram_i* is the information about the educational program of the i-th specialty.

Moreover, the set of subjects, that are provided by the educational program of training specialists in a particular specialty consists of normative and variable components:

$$Course = \{NormCourse_i\}_{i=1}^{N(NS)} \cup \{VarCourse_j\}_{j=1}^{N(VS)}, \quad (2)$$

where *N(NS)* is the number of courses of the normative component of the specialty educational program, *N(VS)* is the number of courses of the variable component of the specialty educational program, and *N(S) = N(NS) + N(VS)*, where *N(S)* is total the number of courses provided by the curriculum of the specialty educational program.

Variable courses of each specialty are courses introduced by higher education institution to better meet the educational and qualification requirements of a person for the

needs of society, more effective use of the educational institution, taking into account regional characteristics, etc. The peculiarity of variable courses is the possibility of in-depth study of individual sections of the course area and taking into account individual professional needs and popular trends in the personal development of educational services consumers.

Thus, the courses selection by the student are a tuple:

$$VarCourse_i = \langle TitleVarCourse_i, CreditECTS, Department/Institution, KnowledgeSkills \rangle, \quad (3)$$

where $TitleVarCourse_i$ is the title of the variable course of the i -th specialty, $CreditECTS$ is the number of credits allocated for course in the curriculum, $Department/Institution$ is the subdivision of the HEI, which provides teaching course, $Knowledge&Skills$ are the knowledge, skills and abilities. Services after studying course.

Moreover, the title of the variable course of the specialty is a set:

$$TitleVarCourse = \{ VarSubjMarkers_k \}_{k=1}^{N(M)}, \quad (4)$$

where $VarSubjMarkers$ is the set of variable subject markers, $N(M)$ is the number of subject markers that characterize the course.

Subject markers are keywords and their combinations that characterize the main subject of study of a particular course. That is, the courses subject markers are keywords that characterize certain sections of courses in the professional field, methods of study and scope (eg, applied grammar, computational linguistics, mobile devices, machine learning, etc.). Therefore, the subject marker has the form:

$$VarSubjMarkers_i = \{ Keywords_j \}_{j=1}^{N(VM)}, \quad (5)$$

where $N(VM)$ is the number of the set of subject markers of the i -th specialty.

To form a contingent of students to study selective courses for the next academic year, departments are introduced to students with a list of selective courses they teach and their annotations, as well as assist students in obtaining full information about the selective course. Analysis of the process of selecting variable courses by students of the Lviv Polytechnic National University for 2018–2020 academic year also shows not very high rates of student activity. 25% -30% of students for 2018–2019 academic year, and 50% of students for 2020 academic year, did not show a desire to independently select courses from a variable block of academic courses (see Fig. 1).

In addition, the name of the course and its short annotation do not fully define the basic knowledge, skills and abilities that the student will receive after studying this course. Since the position of students in obtaining additional information is not always active, when they select guided only by the courses title. That is why most students select courses that contain, popular for this specialty, keywords (eg, “startup”, “hacking”, etc.) or words that characterize the discipline in terms of organizing their own business (eg, “business analysis”, “business management”, etc.).

Although they can get more detailed information about courses by getting acquainted with their work programs at the department that provides teaching, as well as curricula for training specialists in other specialties.

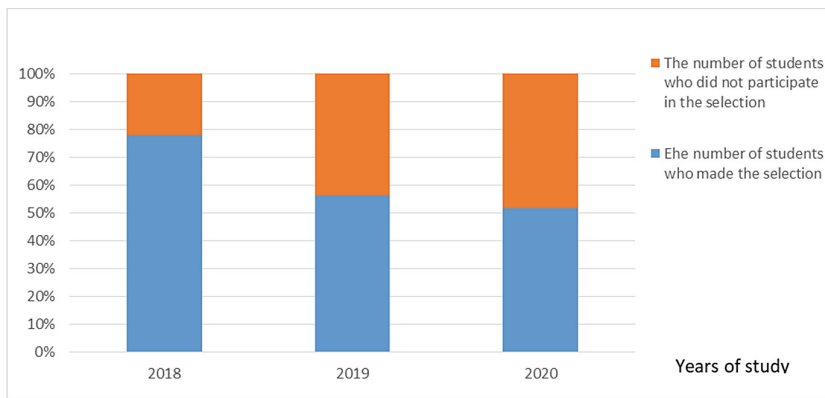


Fig. 1. Analysis of the process of selecting variable courses by students of the Lviv Polytechnic National University for 2018–2020 academic years

Therefore, variable course select of a particular specialty is a set:

$$VarSelect_i = \left\{ \langle TitleVarCourse_j^{(VarSubjMarkers_i)}, \omega(VarSubjMarkers_i)_j \rangle \right\}_{j=1}^{N(VS)} \quad (6)$$

where $\omega(VarSubjMarkers) \in \{0; 1\}$ is a measure of the presence of subject markers in the i -th selected variable course.

The degree of subject markers presence in the i -th selected variable course is a weighting factor that shows the level of probability of selecting a certain course by students of the HEI, taking into account the popularity of subject markers in the course title.

This situation to selection courses is a consequence of the lack of professional experience of students, as well as the peculiarity of their age and psychological development. This should encourage the relevant HEI departments, whose variable courses are not popular for some reason or contain unpopular subject markers, to clarify among students the need for specific areas of development in their professional activities, as well as encourage them to participate in selecting courses as a process. Their formation in the professional sphere and preparation for participation in market relations of the professional sphere.

4 Accordance of Popular Variable Courses to the Main Search Trends

Most consumers of educational services are active users of the Internet. That is why, students get more complete information about the possibilities of applying knowledge, skills and abilities for further development in the professional field from the World Wide Web with the help of search engines.

Each search query, which is part or the entire title of the course, displays only the main part of the information content. That is why the integral display of the user is

formed with the help of related search topics, which specify the scope of professional knowledge, skills and abilities, as well as related topics of information retrieval.

The set of search queries and related topics form a set of keywords. Therefore, a search query is a set:

$$SearchQuery_i = \{Keywords_j\}_{j=1}^{N(Keywords)} \quad (7)$$

where $N(Keywords)$ is the number of keywords that characterize course or professional field of application of subject knowledge, skills and abilities.

For example, for the course “Startup Development Technologies”, related search topics are words/phrases such as “business analysis”, “own business”, “project management”, “information technology”; for the course “Business Planning and Project Management” related search topics are “business”, “projects”, “management”; for the course “Fundamentals of Logistics” related search topics are such as “management”, “economics”, “trade”.

To determine the accordance of variable courses to the trends of search queries, it is necessary to establish a functionality that takes into account the degree of correspondence of the i -th variable course to the subject markers that may be contained in the search query:

$$Adequacy^{(VarSubjMarkers_i)}(SearchQuery)_i \geq \alpha, \quad (8)$$

where $\alpha \in (0,1]$ is the minimum value of the subject marker to match the search query trends. It can be assumed that when such a match is present, the variational discipline will be among the popular ones. Argue that the variable course will be characterized by minimal selection among consumers of educational services or its absence.

5 Results

Analysis of rankings among search queries, according to the service <https://trends.google.com.ua> in 2019–2020 (see Fig. 2) shows trends of growing interest over time in terms related to the organization of their own business. Indicators of interest over time show the popularity of the search term relative to the highest value for a certain period of time. Moreover, 100 is the peak of the term's popularity; 50 means that the popularity of the term is twice less; 0 means that there was not enough data about this term.

These search queries contain subject markers that characterize certain variable courses of the student curriculum.

In particular, the analysis of results of variable course selection by HEI's students on the example of the Lviv Polytechnic National University, shows that most students select courses only by title and were not interested in information content of these courses. The results of variable courses selection by students of the Lviv Polytechnic National University for 2018–2020 academic year, shown in Fig. 3.

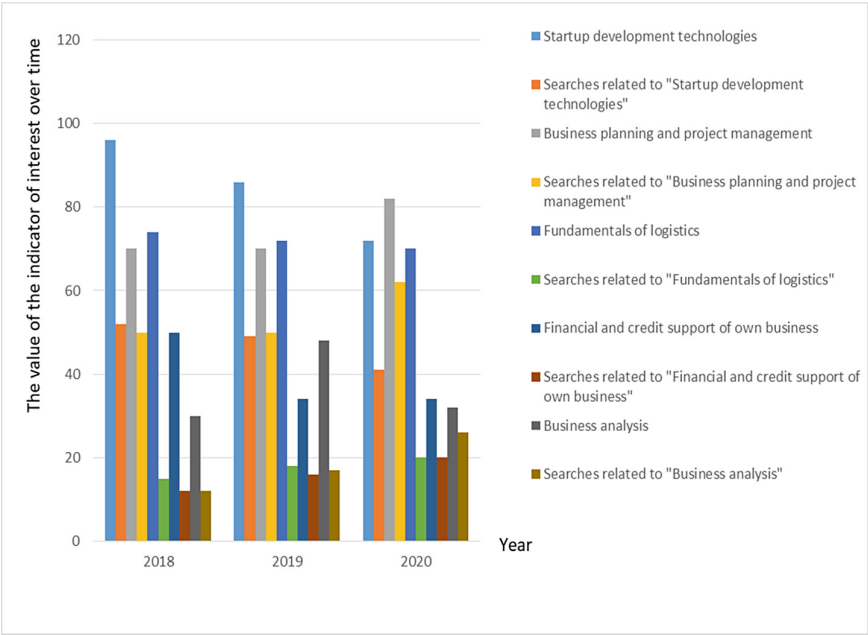


Fig. 2. The value of the indicators of interest over time of popular search queries according to the service <https://trends.google.com.ua/> for the period 2018–2020 academic years

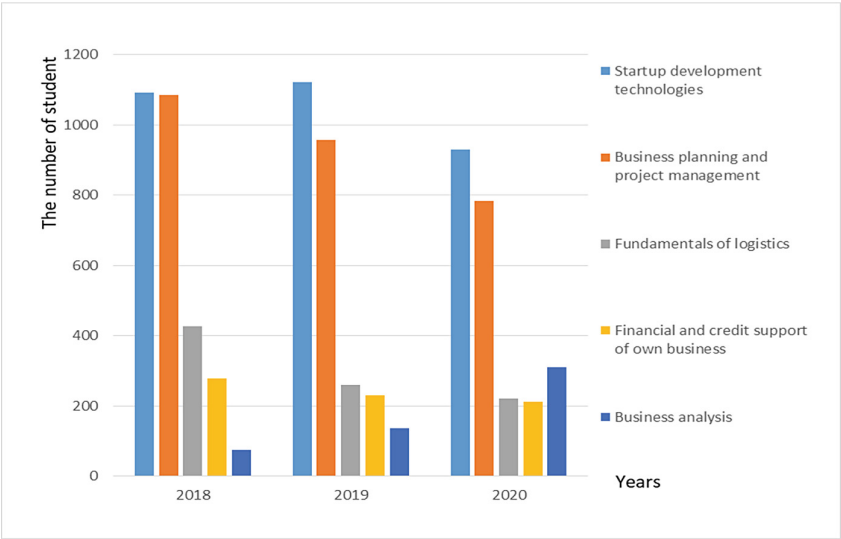


Fig. 3. The results of variable courses selection by students of the Lviv Polytechnic National University for 2018–2020 academic years

Analysis of results of variable courses selection by students of the Lviv Polytechnic National University for 2018–2020 academic years shows that the most popular variable course for three years in a row were “Startup Development Technologies”, “Business Planning and Project Management”, and “Fundamentals of Logistics”. Other “courses-leaders” are a number of economic courses that characterize the course in terms of organizing their own business: “Financial and credit support of their own business”, “Business Analysis”.

Thus, there is a dependence of trends in the variable course selection on the indicator of interest over time of the search query in search engines. And there is a tendency to increase interest over time in related search topics. This is due to the experience of applying certain professional knowledge, skills and abilities in various fields of activity.

Analysis of popular search queries and variable disciplines allows you to form a set of subject markers that characterize the popular variable course (Table 1).

Table 1. Set subject markers of popular variable courses

The title of a variable course	Subject markers	Related subject markers
Startup development technologies	Startup, technology, development,	Information technology, artificial intelligence, business, project, capital,
Business planning and project management	Business planning, project, management	Management, organization, business plan, planning
Fundamentals of logistics	Logistics	Transport, management, manager, system
Fundamentals of logistics Financial and credit support of own business Business analysis	Finance, credit, own business	Business plan, business, project, management
Business analysis	Business, analysis	Business plan, project, organization

Analysis of search queries that do not have search activity is shown in Fig. 4

Analysis of the rejection of variable courses by students of the Lviv Polytechnic National University are “The role of religion”, “Business career” (2018 academic year), “Sociology of conflict”, “Environmental marketing”, “Methods and models of the transport system” (2019 academic year), “Data Visualization”, “Infographics” (2020 academic year). Analysis of the rejection of variable course by students of the Lviv Polytechnic National University for 2018–2020 academic year shown in Fig. 5.

Analysis of the results of the deviation of the course selection and search queries containing subject markers of these courses shows the lack of unambiguous relationship between them. Among the rejected courses are those that are characterized:

a steady increase in activity in search queries and related queries (eg, “Religion”, “Sociology of Conflict”). Given the inertia of education, one can expect to select these

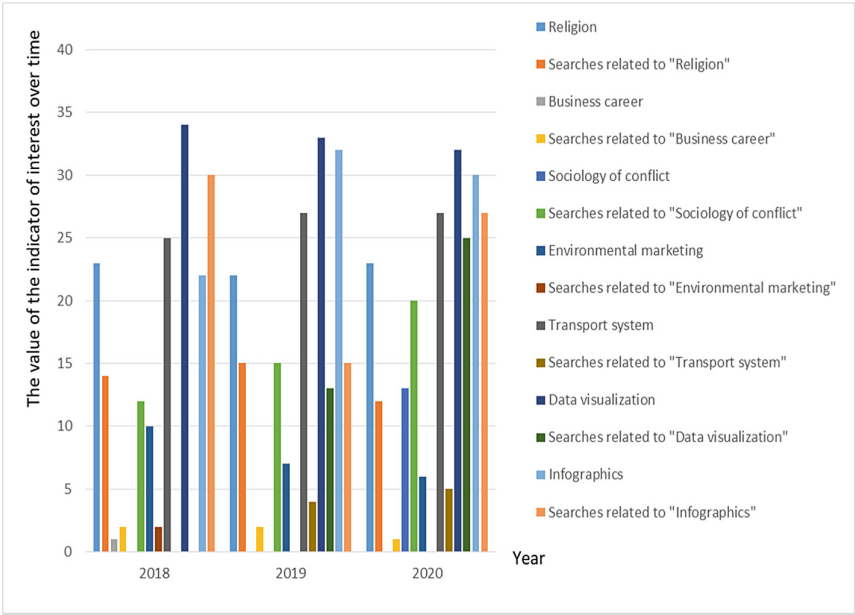


Fig. 4. The value of the indicators of interest over time unpopular search queries according to the service <https://trends.google.com.ua/> for the period 2018–2020 academic years

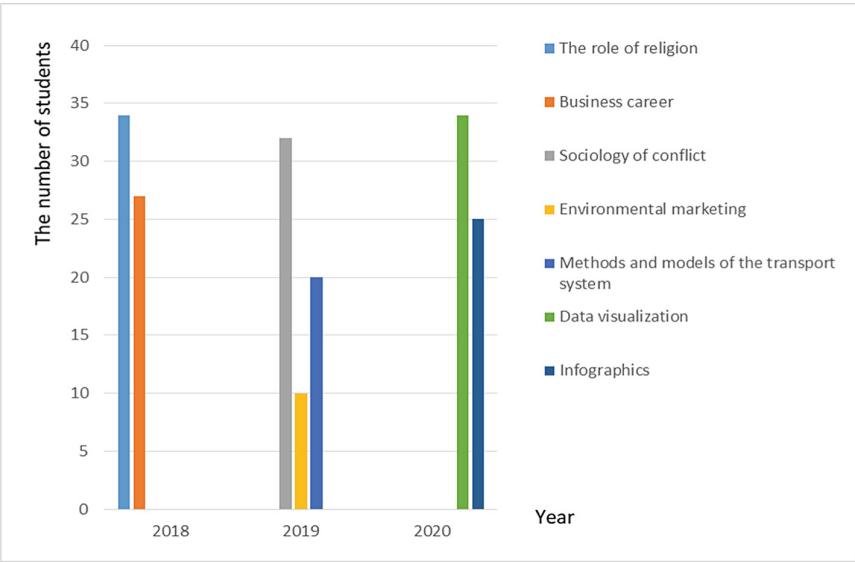


Fig. 5. The results of the rejection of variable course by students of the Lviv Polytechnic National University for 2018–2020 academic year.

courses, but while maintaining or improving search query trends and related search topics. That is, students can come to this selection in a certain time;

non-uniform increase in activity in search queries and related queries (eg, "Methods and models of the transport system", "Data visualization", "Infographics"). This indicates that students' preferences are volatile. Therefore, it is necessary to carry out explanatory work among consumers of educational services and adapt the content of the training course to labor market trends.

a steady decline in activity in search queries and related queries (eg, "Business Career", "Environmental Marketing"). The result of such a deviation may be "not interesting" for consumers of educational services, the title of the course. That is why it is necessary to adapt the title and information content of the course to the trends of the labor market and popular areas of youth activity.

Thus, the popularity of selected variable courses is closely relate to popular trends in the labor market and personal development. That is why, taking into account, the trends of search queries allow us to predict the behavior of consumers of educational services and predict the enrollment of students for elective subjects according to the educational program of the specialty.

6 Conclusion

Thus, the linguistic analysis of results of variable disciplines selection by HEI's students shows that the students' selection does not always take place in terms of the need for appropriate knowledge, skills and abilities for their professional development. For the most part, students selection from new trends in certain professions, identifying relevant subject markers in the title of the course, but usually not interested in the availability of appropriate information content of courses in accordance with the professional needs of consumers of educational services. The correspondence between popular courses and search queries was established with the help of appropriate subject markers is a result of linguistic analysis of variable courses selection. It was also established that among the reasons for the deviation of some courses are the trends of increasing and decreasing search activity of Internet users using subject markers of these courses. Linguistic analysis of variable course selection by higher education seekers also allows HEIs to effectively plan the provision of educational services and remain competitive in the market of educational services.

References

1. Kulik, O.: Choice of system of indicators for estimation of quality of grant of educational services by educational establishments. *Sci. J. "ScienceRise"* **7/1**(12), 47–53 (2015)
2. Schudlo, S.: Mechanisms of forming of social conventions are in higher education as condition of providing of her quality. *S.P.A.C.E.* **1**, 31–37 (2016)
3. Alawamleh, H.S., Bdah, A., Alahmad, N.: The impact of planning on the quality of educational programs at AlBalqa' applied university. *Int. J. Bus. Administ.* **4**(5), 8–50 (2013)

4. Watty, K.: Quality in Accounting Education: What Say the Academics? *Qual. Assur. Educ.* **2**(13), 120–132 (2005)
5. Mace, J.: Higher education and business. *Higher Educ. Rev.* **7**(25), 68–72 (2008)
6. Mikhno, N., Sorokina, L.: Trend analysis of the educational choice of university entrants as tool to improve the quality of educational service. *Int. J. Inf. Commun. Technol. Educ.* **6**(3), 36–39 (2017)
7. Creed, P.A., Patton, W.: Differences in career attitude and career knowledge for high school students with and without paid work experience. *Int. J. Educ. Vocat. Guidance* **3**, 21–33 (2003)
8. Kudirko, O.: The choice of profession is choice of course of life: agitation work with potential entrants. In: *Studentcentrism in the System of Providing of Quality of Education in an Economic University*, pp. 99–100 (2016)
9. Kniveton, B.H.: The influences and motivations on which students base their choice of career. *Res. Educ.* **72**, 47–57 (2004)
10. Miller, L., Lietz, P., Kotte, D.: On decreasing gender differences and attitudinal changes: factors influencing Australian and English pupils' choice of a career in science. *Psychol. Evol. Gender* **4**(1), 69–92 (2002)
11. Small, J., McClean, M.: Factors impacting on the choice of entrepreneurship as a career by Barbadian youth: a preliminary assessment. *J. Eastern Caribbean Stud.* **27**(4), 30–54 (2002)
12. Bailyn, L.: Academic careers and gender equity: Lessons learned from MIT. *Gender Work and Organisation* **10**(2), 137–53 (2003)
13. Shilinh, A., Zhezhnych, P.: Linguistic approaches to the planning of educational services in higher education institution. *ECONTECHMOD* **7**(4), 13–20 (2018)
14. Zhezhnych, P., Shilinh, A., Melnyk, V.: Linguistic analysis of user motivations of information content for university entrant's web-forum. *Int. J. Comput.* **18**(1), 67–74 (2019)
15. Chyrun, L., Gozhyj, A., Yevseyeva, I., Dosyn, D., Tyhonov, V., Zakharchuk, M.: Web content monitoring system development. In: *CEUR Workshop Proceedings*, vol. 2362, pp. 126–142 (2019)
16. J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, Y. Burov, "Information resources processing using linguistic analysis of textual content", *Proceedings of the 2017 IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2017*, 2(8095038), pp. 573–578, 2017.
17. Sebastiani, F.: Machine learning in automated text categorization. *ACM Comput.* **34**(1), 1–47 (2002)
18. Bisikalo, O.V., Vysotska, V.A.: Identifying keywords on the basis of content monitoring method in Ukrainian texts. *Radio Electron. Comput. Sci. Control Zaporizhzhya Natl. Tech. Univ.* **1**, 74–83 (2016)
19. Lytvyn, V., Vysotska, V., Veres, O., Rishnyak, I., Rishnyak, H.: Classification methods of text documents using ontology based approach. In: *Advances in Intelligent Systems and Computing*, vol. 512, pp. 229–240. Springer (2017)
20. Mukalov, P., Zelinskyi, O., Levkovich, Tamavskiy, P., Pylyp, A., Shakhovska, N.: Development of system for auto-tagging articles, based on neural network. In: *CEUR Workshop Proceedings*, vol. 2362, pp. 116–125 (2019)
21. Paramonov, I., Lagutina, K., Mamedov, E., Lagutina, N.: Thesaurus-based method of increasing text-via-keyphrase graph connectivity during keyphrase extraction for e-Tourism applications. In: Ngonga Ngomo, A.C., Křemen, P. (eds.) *KESW 2016. Communications in Computer and Information Science*, 649, pp. 129–141. Springer, Cham (2016)
22. Lytvyn, V., Vysotska, V., Burov, Y., Veres, O., Rishnyak, I.: The contextual search method based on domain thesaurus. In: *Advances in Intelligent Systems and Computing*, vol. 689, pp. 310–319 (2018)



Poetic Discourse Processing Applying Graph Theoretic Rules

Olena Flys^(✉) and Maria Bekhta-Hamanchuk

Applied Linguistics Department, Lviv Polytechnic National University, Lviv, Ukraine
olenkabondaruk90@gmail.com, mariabekhta@gmail.com

Abstract. In this article, we present a way of graph theory application to analyze the structure and content of poetic discourse. The study has been carried out in terms of applied linguistics and illustrates the integration of information systems technologies. The main contribution is an attempt to implement the open-source text analytics software to analyze the poetic discourse of the late 16th – early 17th centuries and provide its infographics visualization. The topicality of graph theory and data visualization applications in linguistic studies are connected with the ability to structure the retrieved information, identify the contextual relations between the key phrases and language units, and make predictions regarding the dynamics of discourse organization. Overall, the research findings indicate the most frequent connections between the lingual cultural type of woman and the lexical variables used to depict physical, social, or spiritual objects of the reality in English culture.

Keywords: Cognitive approach · Discourse · Graph theory · Poetic discourse · Visualization

1 Introduction

Modern tendencies of increasing the amount of information and raising the necessity to take into account a large number of interrelated factors require both improving the quality and speed of existing information systems. The most efficient way of solving these issues is the use of modern control technologies. As a result, the graph construction methods are regularly used to solve the problems of data analysis, presentation, and processing.

In our previous researches, we have already designated the essential role of graph theory regarding applied linguistic researches and revealed the terminology system in the correlation of graph theory and linguistic studies [1]. This paper attempts to present the possibilities of investigating the poetic discourse structure and content using the rules of graph theory and visualization applications. The structure of both discourse and language is of high interest for modern scientists, especially when it comes to the applied linguistic tasks such as automatic text processing, computer modeling problems, and the graphic text construction [2–4].

The analysis of complex text systems requires the approaches that do not concentrate as much on the accuracy and strictness of the quantitative methods, but allow the

fuzziness and partial units. That is why the graph theory is often applied when the initial information has semantic nature and is presented either in the form of interconnected system elements or in the form of a weighted graph that connects the elements of this system [5]. The development of graph theory methodology is aimed at revealing the new opportunities that can improve the efficiency of information processing.

2 Introduction to Discourse

According to the recent applied linguistic studies, discourse analysis has become one of the significant issues in the modern termbase. It might be explained by the fact that discourse contains not only the linguistic components (language units), but also the extralinguistic elements (cognitive structures, knowledge of the world, person's worldview, ideological content, set of frames, and graph structures) [6;8; 7].

A lot of attention is devoted to the discourse as a coherent text that contains extralinguistic, pragmatic, paralinguistic, sociocultural, psychological and other factors. It is defined as "a conceptual text" that participates in the interaction of people and their consciousness [8:136]. The researchers revealed that discourse can be presented either in the form of a whole text or in its smaller fragments that are surely "united by a coherent logical structure and language relations aimed at implementing some author's informative or social perspectives" [9:92]. The perception of discourse as a sequence of coordinated components requires the interpretation of the bonds between them [6:35] that demands additional extralinguistic knowledge from the interpreter.

Scientists that are engaged in researching discourse nature emphasize that it is a complex communicative unit. Some of them define this term as "the sequence of interrelated statements" that are connected with common objectives [10], while others believe that discourse is "a text of coherent speech" that consists of the determined sequence of lexical units characterized by semantic relations [11:13].

Considering the variety of interpretations that we found in the process of analyzing the term "discourse" it should be noted that the common idea among the researches is the recognition of the complex nature of discourse. On the one hand, it is an activity and a process, but on the other hand, it is a text and a product that is characterized by a number of extralinguistic parameters. Due to it, the interpretation of poetic discourse requires to include all the above-mentioned approaches to discourse understanding.

3 Cognitive Approach to Poetic Discourse Analysis

Over the past two centuries we see the rising interest to poetic discourse analysis that has even become an issue of structural (A. A. Kaminchuk, A. G. Revzin) and graphical stylistic (P. O. Kovalev, M. V. Panov, D. Peddison) approaches to discourse structure researches.

In this study, we are aimed to reveal the cognitive approach to poetic discourse as the creation of graphs implies the construction of the cognitive model [1:610]. According to this approach, language is "an instrument of organizing, generating, and transmitting the information" [12:114], it is "a cognitive mechanism, which develops a system of signs that determine how to represent and transform the information" [14:53]. The most

important thing about this approach is “the systematic description and explanation of the mechanisms people use to acquire the language” [15:21]. In other words, the main function of this approach is to clarify the way knowledge is presented and what procedure should be used to process it.

According to the cognitive approach, poetic discourse is interpreted as a cognitive model of archaic thinking that depends on the author’s individual world perception. Poetic discourse is a set of frames that has “a hierarchical system of semantic relationships in the text that serves to reflect language and conceptual poet’s world view” [16:3–15]. In this research, we are to deal with the organized system of poetic texts consisting of figurative and lingual elements that contain both pragmatic and sociocultural interrelated features [17].

The followers of the cognitive approach believe that the discourse formation is closely related to the processes of “verbal behavior establishment” [19:19] and consider it as a cognitive process. Basing on mental structures and individual knowledge, the representatives of this approach are aimed to explain the author’s motives when they made a choice of lexical units to explain or describe particular facts. Poetic discourse has such peculiar features such as human-centricity, culturalism, and spirituality that are manifested not only in the use of linguistic means but also through its structure – rhyme, rhythm, graphics, strophic.

4 Discourse Structure Construction

One of the most authoritative theory that describes the discourse structure is the Rhetorical Structure Theory proposed by W. Mann and S. Thompson. The researchers distinguish three levels in the discourse structure: global, local, and syntactic. The Rhetorical Structure Theory is based on the fact that every single discourse unit is connected to at least one other unit of the same discourse through some meaningful connection and semantic relations. They believe that discourse has a tree structure [19]. The discursive units of rhetorical relations can be of various sizes from maximum (the direct components of the whole discourse) to minimal (the individual predications). The discourse structure is hierarchical and all its levels obtain equal following rhetorical relations: sequence, volitional result, non-volitional result, condition, concession, joint, elaboration, background, purpose, otherwise, etc. [20:257].

The global structure of discourse indicates its division into large components: episodes in a story, paragraphs in a newspaper article, etc. The written discourse is characterized by the visual method of marking the global structure that is a graphic paragraph. It is generally assumed that a paragraph boundary is a change of topic [21:95] or a connectivity decrease. Connectivity is usually treated as a combination of characteristics common to a certain discourse fragment. For example, T. Givón distinguishes between referential, temporal, spatial, and eventual connectivity [22].

In contrast to the global structure, the local structure of discourse describes its division into minimal components that refer to the discursive level. Most of the modern approaches consider clauses to be these minimal components [23].

The structure of poetic discourse is believed to be totally different from the other common types of discourse because of their structure peculiarities – rhythm, syllables' combinations, and inflexion. This type of discourse is also characterized by the specific rules of grammar and various poetic conventions such as meter (iambic, trochaic, pyrrhic, etc.), feet (dimeter, hexameter, etc.), stanza counts, line counts, rhyming rules, and stress sequences.

Poetic discourse is considered to have a specific hierarchical structure, where nodes are connected with the rhetorical relations. Each node of the rhetorical hierarchical structure is further realized as a set of semantic units – lexical or grammatical. Most of these rhetorical relationships are asymmetrical and binary, also they always contain a nucleus and a satellite.

5 Discourse Processing and Graph Theory

The benefits of applying graph theoretic rules and formulas to poetical discourse are the ability to structure poems on “topics or subjects that traditional poets might ignore, yet for which an audience might exist” [24].

Defining discourse as “a set of binary relations between a current utterance and the preceding discourse” [25] illustrates its close relation to the graph construction schemas. Using graph theory formulas and applications, we may represent rhymes as nodes in a directed graph, or a simple graph. A line, that we get, connects two words indicating that they are linked as rhymes according to the rhyme scheme of the poem. We may say that a graph is a network with words embedded in it. This graph may be explored and interpreted using the results from graph theory to provide both phonological and semantic insights [26]. Taking into account the main graph theoretical postulates [27] we may distinguish the poem as a graph $GG = (VV, EE)$ that consists of nodes (vertexes) as the lexical units and contextual relations among them as the arcs.

Using graph theory to analyze the structure of discourse illustrates the syntax relations between the predication in discourse. It is not crucial for graphs construction whether the presented relation is expressed by a union of the corresponding semantics, a comma, or whether it combines the independent sentences or groups of sentences. Graph theory is characterized by a formalism that foster to represent discourse in the form of networks of discursive units and rhetorical relations.

The main importance of using graph theory in linguistic researches is the possibility to provide alternative interpretations of the same text. In other words, more than one graph of a rhetorical structure can be constructed for the same text. However, this multiplicity is limited. There is a number of weighty confirmations that the functionality of graph theory largely models reality and represents the high importance in revealing how the discourse works in real life. Because of some rules, many satellites in the rhetorical pairs can be omitted, while the resulting text remains coherent and quite representative in relation to the source text.

The basic tasks that we can fulfill with the help of graph construction might be divided into the following categories such as cognitive analysis problems, graph decomposition into the subgraphs, aggregation, or generalization (e.g. clustering, classification), graph vertex allocation (e. g. linear, plane, volumetric), graph vertex covering (e. g. the partition into tolerance or similarity classes), graph coloring, graph packing, graph cycles indication, isomorphism determination, isomorphic embedding, isomorphic graph intersection, and determining graph similarity [28].

The main structural features of the text are integrity and connectivity that can reflect its substantial and structural essence. While analyzing a text, it is crucial to select the segments that correspond to the isolated, pendant, and deadlock vertex of a graph. The isolated vertexes are not incident to any of the edges (arcs) of a graph, which means that this segment of a graph is not connected with the other segments. The pendant vertexes correspond to the segments that cannot be reached from the other segments, while the deadlock vertex illustrates that you cannot reach the other segments of a text from a certain segment.

You can find the isolated, pendant and deadlock vertexes of a graph with the help of adjacency matrix of $A = \|a_{ij}\|$ graph, where every vertex k ($k = (1, n)$ -, n is the number of vertexes in a graph) is designated with a vector $a(k) = (a_k, a_k)$ that has the following components:

$$a_k = \sum_{j=1}^n a_{kj}, a^k = \sum_{i=1}^n a_{ij}, \quad (1)$$

where a_k is the sum of k -row elements and a^k is the sum of k -columns of the adjacency matrix. The value a_k is determined by the number of arcs emerging from the vertex k , while a^k is the number of arcs that enter into it. If $a_k = a^k = 0$, the vertex k will be isolated; if $a_k = 0$, it will be deadlock; if $a^k = 0$, it will be pendant [29].

The presence of the isolated vertexes in a graph mostly indicates the disconnectedness (lack of integrity) of the text. The pendant vertexes must correspond to the final positions of the text, while the deadlock vertexes must be presented by a segment that corresponds to the central idea of the text.

The analysis of text segments' relations is primarily aimed at revealing the loops, contours, and strongly connected subgraphs in the corresponding graph. We interpret the loop as the existence of a certain connection between the input and output peculiar to the same segment, i. e. the insularity of argumentations in this segment. The contour is responsible for constructing a path, that is an alternating sequence of edges (arcs) and vertexes where the initial and final vertexes always coincide. Taking into consideration these statements, we can emphasize the absence of causal relationships in the text.

6 Graph Theory Application and Visualization

In this research, we decided to apply the open-source text analytics software KH Coder (Version 3, 2016) [30] in order to reveal the content structure of poetic discourse and its graph construction peculiarities. Mostly, this application is used to provide the quantitative content analysis, text mining tasks, and corpus linguistics experiments. It can process texts in English, French, German, Italian, Portuguese, Spanish, and Japanese languages.

This platform is able to accomplish the diversity of tasks with the input raw texts, such as KWIC search and statistics analysis, collocation statistics, occurrence network creation, word map construction, multidimensional scaling, cluster analysis, and correspondence analysis. KH Coder provides various types of search and statistical analysis functions using the backend tools such as Stanford POS Tagger, Snowball stemmer, MySQL, and R [31].

In this research, we apply the functions of KH Coder software to analyze the lingual corpus that consists of 1692 poetic texts written during late 16 – early 17 century in England (see Fig. 1).

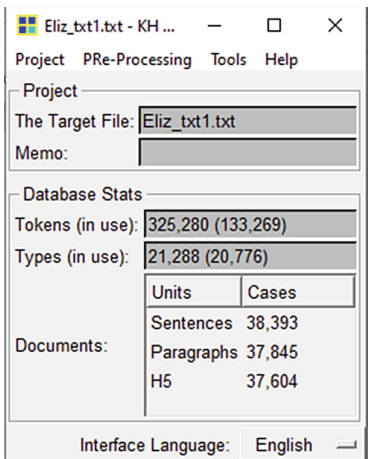


Fig. 1. KH Coder software processing of poetic discourse

At the beginning, we used the command “Frequency list” that enables a user to see 100 most often used lexical unit (see Fig. 2, 3) by counting the base forms and lemmas. According to the obtained results, we may see that the most often used units are the following nouns “love”, “eye”, “heart”, “man”, “life”, “beauty”, “day”; verbs “make”, “doth” (=“do”), “let”, “come”, “love”, “say”, “hath” (=“has”); adjectives “sweet”, “fair” and the pronoun “thy” / “thee” (=“you”). All this indicates, that the poetic discourse of the analyzed historical period put a human into the center of the world-centricity and was full of spiritual, esthetical, and philosophical ideas.

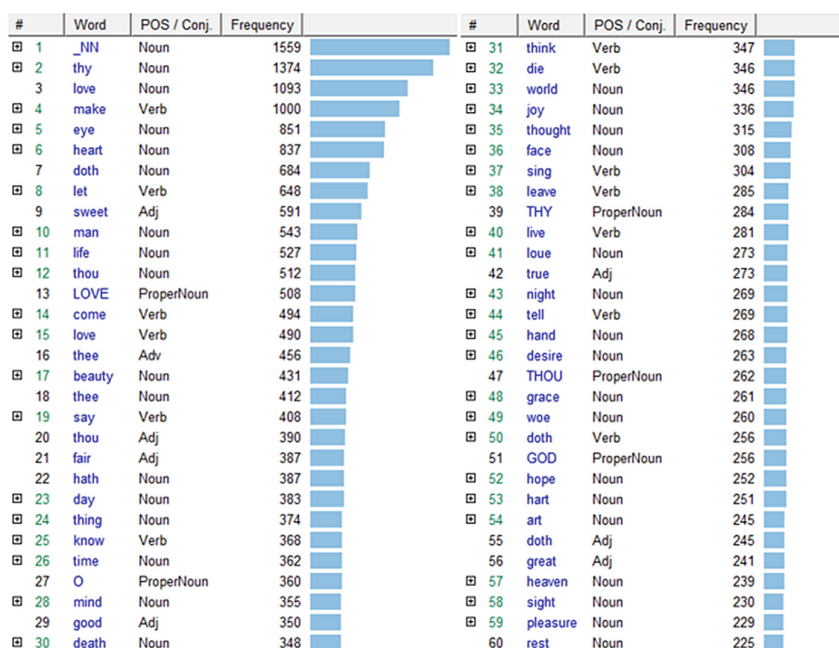


Fig. 2. The results of “Frequency list” command application (Words 1–60)

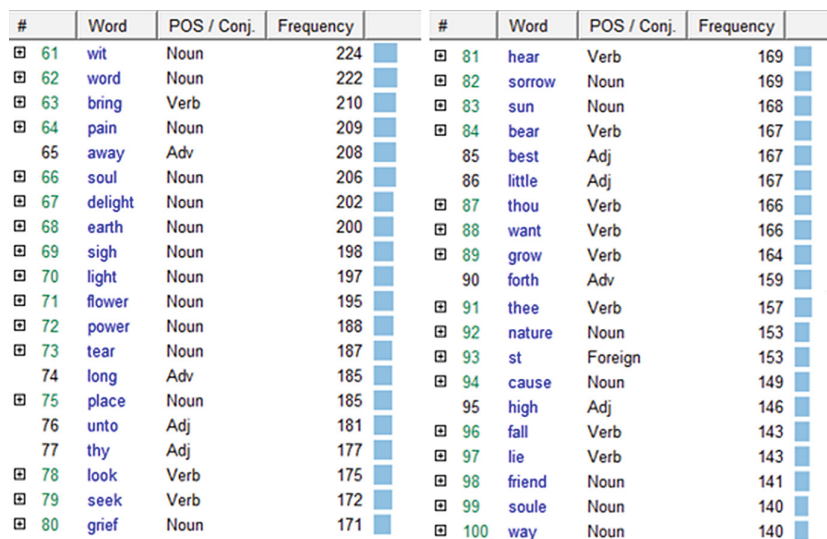


Fig. 3. The results of “Frequency list” command application (Words 61–100)

The next function that we applied to the researched corpora was “KWIC Concordance”. It helps to see the way the extracted word is used in the target text (see Fig. 4).

The search results are sorted in the following orders: 1) according to the condition specified in “Sort 1” condition; 2) the items in the same set are resorted according to the “Sort 2” condition; 3) the items still in the same set are resorted according to the “Sort 3” condition [31:37].

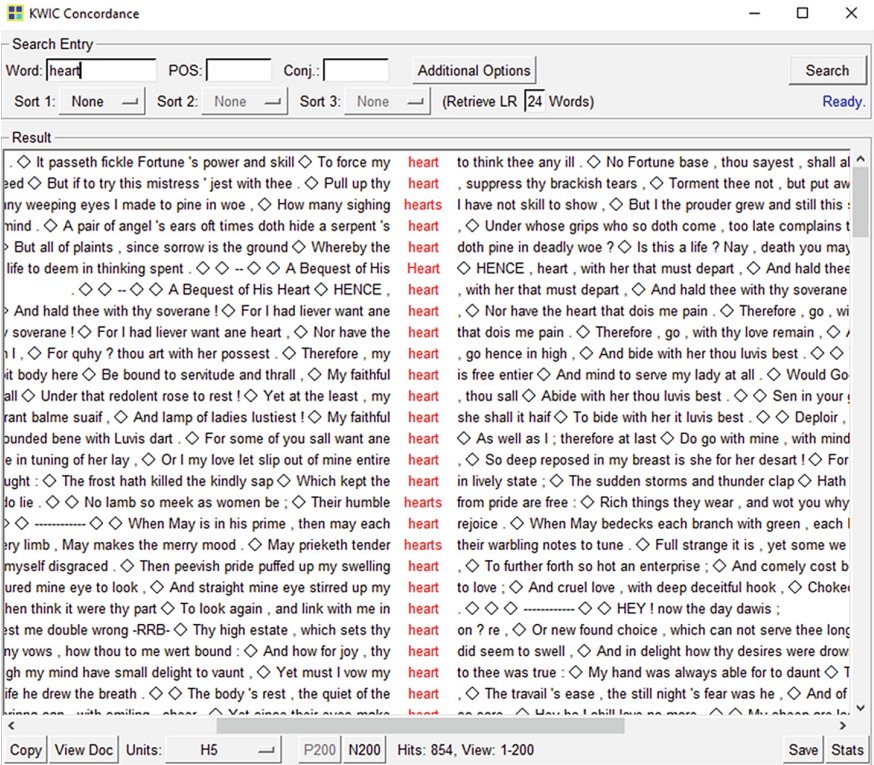


Fig. 4. The results of “KWIC Concordance” command application (“LOVE”)

After the system receives the input text, it proceeds to the analysis of the relevance bonds between the text units by using the “fuzzy” method. Primarily, it determines all the relationships of this sort that exist between the key text phrases or lexical units, only after that the system can provide the visualization of a directed graph that reflects these relations. The graph we receive, as well as the results of its structure analysis, illustrates the other function of KH Coder – “Co-Occurrence Network”.

The retrieved graph (see Fig. 5, 6) allows us to determine the most common key phrases in the analyzed poetry texts that correspond to the largest graph vertexes, as well to observe the structure of the relationships between them: LOVE (sweet love, do love, true love, love life, thee love), THY (thy make, thy eye, thy beauty), HEART (you heart, make heart, hear – eye), LIFE (life – die, life – death, long life, life – joy, do life, live life), EYE (face – eye, fair eye, let eye, thy eye).

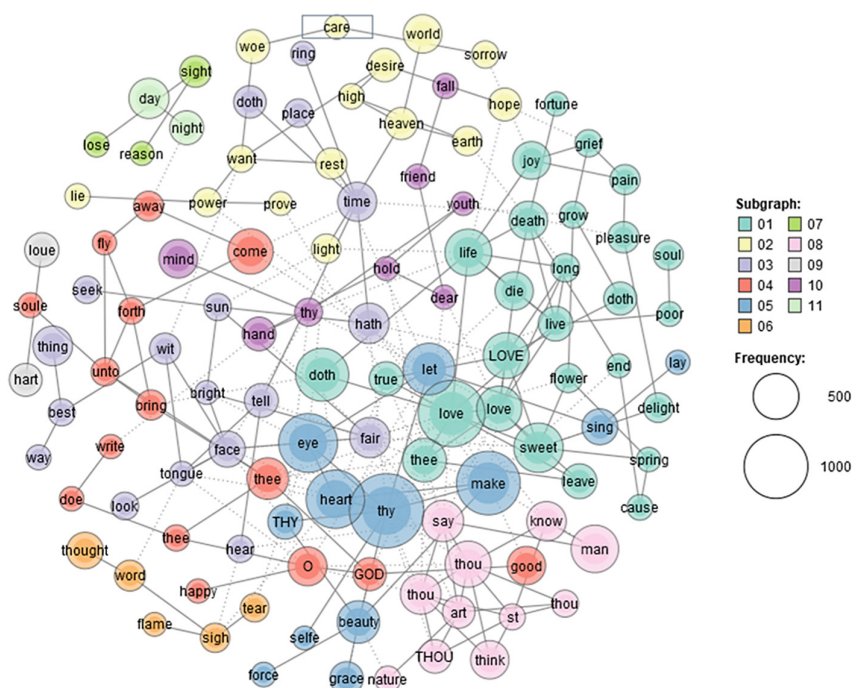


Fig. 5. Graph retrieved after the “Co-Occurrence Network” command application

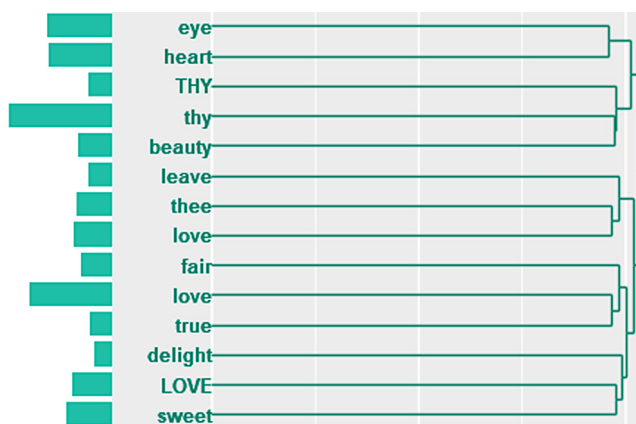


Fig. 6. Most common key text phrases and lexical units

Additionally, we decided to apply the supplementary tool for infographics visualization – Wordle cloud program (Version 0.2) [32]. It allows us to conduct a specified analysis of the target text, create the infographics from it, establish the links between words, and identify the keywords. Wordle tool provides an ability to create the “word

2. Khomytska, I., Teslyuk, V.: Authorship and style attribution by statistical methods of style differentiation on the phonological level. In: *Advances in Intelligent Systems and Computing III. AISC 871 (Selected Papers from the International Conference on Computer Science and Information Technologies, CSIT 2018, September 11–14, Lviv, Ukraine)*, Springer (2019)
3. Dilai, M., Levchenko, O.: Discourses surrounding feminism in Ukraine: a sentiment analysis of Twitter data. *Advances in Intelligent Systems and Computing III. AISC 871, Selected Papers from the International Conference on Computer Science and Information Technologies, CSIT 2018, September 11–14, Lviv, Ukraine*, 47–50. Springer (2018)
4. Perkhach, R.-Y., Shyika, J.: The methodology of frequency dictionaries to the instructions to the medical products. In: *Advances in Intelligent Systems and Computing III. AISC 871, Selected Papers from the International Conference on Computer Science and Information Technologies, CSIT 2018, September 11–14, Lviv*, 26–30. Ukraine, Springer (2019)
5. Vagin, V.N.: *Deduction and generalization in decision making systems*. Nauka, Moscow (1988)
6. Van Dijk, T.A.: *Language. Cognition. Communication*, Progress, Moscow (1989)
7. Langacker, R.: Discourse in cognitive grammar. *Cogn. Linguist.* **12**, 143–188 (2001)
8. Arutyunova, N.D.: Discourse. *Linguistic Encyclopedic Dictionary*, pp. 136–137. Soviet Encyclopedia, Moscow (1990)
9. Kusko, K.Ya.: Cognitive-discursive potential of informative transfer. *Bulletin of Kharkivskyi nat. un-ty. by V.N. Karazin*, 2004, No. 635, pp. 91–93 (2004)
10. Paddison, G.A.: *Discourse on Poetry*. Nabu Press, Boston (2011)
11. Borbotko, V.G.: *General theory of discourse (principles of formation and generation of meaning)*. Kubansk state un-ty, Krasnodar (1998)
12. Flys, O.: Woman lingual cultural type analysis using cognitive modeling and graph theory. In: *Advances in Intelligent Systems and Computing (AISC), Selected papers from the International Conference on Computer Science and Information Technologies, CSIT 2019, Lviv, Ukraine*, vol. 1080, pp. 602–619 (2019)
13. Ivanchenko, V.O.: Psychosomatic component nomination features of French enological culture. *Linguist. Concept. Pictures World* **43**(part 2) (2013)
14. Kubryakova, E.S., Demyankov, V.Z., Pankrats, Yu.G., Luzina, L.G.: *Brief dictionary of cognitive terms*. Philological Faculty of Moscow State University by M. V. Lomonosov, Moscow (1996)
15. Demyankov, V.Z.: Cognitive linguistics as a kind of interpretive approach. *Quest. Linguist.* **4** (1994)
16. Skripnik, A.V.: *French poetic discourse of the Middle Ages: dissertation abstract*. Kyiv National University by T. Shevchenko, Kyiv (2005)
17. Pilar, A., Hernández, R.: Constructing meaning through creative categorial extension in poetic discourse. *CoDiS Working Papers N.1, Universidad de Salamanca*, pp. 1–22 (2012)
18. Kubryakova, E.S.: *Nominative aspect of speech activity*. Nauka, Moscow (1986)
19. Mann, W.C.: *Rhetorical Structure Theory: A Theory of Text Organization*. ISI: Information Sciences Institute, Los Angeles, CA, pp. 1–81 (1987)
20. Mann, W.: Rhetorical structure theory: toward a functional theory of text organization. *Text* **8**, 243–281 (1988)
21. Brown, G., George, Y.: *Discourse Analysis*. Cambridge University Press, Cambridge (1983)
22. Givón, T.: *Syntax: A Functional-Typological Introduction*, vol. 2. Benjamins, Amsterdam (1990)
23. Testeleys Ya, G.: *Introduction to the General Syntax*. RHHU, Moscow (2001)
24. Parker, P.M.: *An Introduction to Graph Theoretic Poetry*. Access mode (2010). <https://www.totopoetry.com/credits/graphtheoreticpoetry.htm>
25. Hobbs, J.R.: Coherence and Coreference. *Cogn. Sci.* **3** (part 1), 67–90 (1979)

26. James, J.: Re-Weaving the Word-Web: Graph Theory and Rhymes. PBL5 5: General Session, pp. 129–141 (1979)
27. Biggs, N., Lloyd, E., Wilson, R.: Graph Theory. Oxford University Press, Oxford (1986)
28. Bershtein, L.S., Karelin, V.P., Tselykh, A.N.: Models and decision-making methods in integrated intelligent systems. Rostov (1999)
29. Ganicheva, A.V., Ganichev, A.V.: Graph method of texts analysis. World Linguist. Commun. **4**(46), 66–73 (2016)
30. KH Coder platform <https://kxcoder.net/en/> Accessed 15 July 2020
31. Higuchi, K.: KH Coder 3 reference Manual (Version 3.Beta.01a). Japan (2016)
32. Wordle platform. <https://www.wordle.net/>. Accessed 15 July 2020



Textual Features and Semantic Analysis of the Reddit News Posts

Solomiia Albota^(✉) 

Lviv Polytechnic National University, Bandera Street, 12 79013 Lviv, Ukraine
Solomiia.M.Albota@lpnu.ua

Abstract. The phenomenon of social networking platform has been considered in the article. The Reddit online social chatting community has been selected as a basis for the study. The features of linguistic, psychological, textual semantic analysis have been applied to ten discussion comment charts concerning the latest random news – the death of Indian actor, Irfan Khan. The comments in Reddit social network have been allocated and linguistically marked with implicit or explicit comments. Linguistic Inquiry and Word Count (LIWC) program has been used. It proved the textual analysis of the Reddit comment section statistically. The semantic analysis results of both previous papers concerning the notion of conflict situation and contradictions within the social media have been compared.

Keywords: Reddit community driven discussion platform · Conflict situation · Comment · Linguistic markers · Linguistic inquiry and word count (LIWC) · Textual features · Semantic analysis · Reddit news posts · Comparison

1 Introduction

Information technology has advanced over the last decade and it has been discovered that people tend to communicate with each other using social communication channels. Social networks, messengers, media have spread among people all over the world. As it has been reported, a majority of the world's population uses the internet, and more than 80% of Internet users are involved in the developed countries [1]. Estimating calculations, every minute nearly 3 million posts are shared on Facebook, 350 000 posts are twitted, 4 million queries are submitted in Google [2, 3]. Social networks facilitate communication process, thereby encouraging computer-driven research on personality behavior patterns. The interconnection between language and psychological personality features has been greatly studied in psycholinguistics. Based on the language a person uses, it is often enough to judge on his/her intentions, emotional state and dominant cognitive processes. Today, online social networks and their role in human development face an immense challenge in terms of their relevance in interdisciplinary research.

If we talk about social communities and their online communication, textual space is of relevance. Usually, any kind of discussion within social networks has its contradictory nature, as the human being tends to have conflicts while communicating. Conflicts during conversation may start at its beginning, climax or ending. Even the whole conversation

may involve contradictions and arguments. The first and foremost, people should be aware of how to cope with one's sharp opinions, contradictory statements or arrogant nature. Of course, it is not forbidden to reply in any manner people used to do, but still one should keep in mind that respect and rapport are the key to productive conversation. There are also some common scenarios when interlocutors pretend not to reply or ignore some responses. In a psycholinguistic way, it is allowed and can be interpreted in various approaches, but in terms of conflict situations it will be regarded as an unresolved contradictory situation or a conversation with a dead end.

2 Studies Referring to the Issue

Social network platforms have been regarded as an information source, sharing channel, immediate connection tool [4]. The participants of certain social community gain lots of significant and useful information. Moreover, nowadays there is a range of preferences in internet social environment, and every user may determine which one to choose and for what purpose it will serve.

One of such online chatting communities [5–8], where one can find and share the cutting-edge data (e.g. recent news, opinions on the following topics), is Reddit (<http://www.reddit.com>). It is an internet social network involving news forum, the structure of which has a specific layout [9, 10]. In general, it should be stated that Reddit consists of subreddits (areas of topical interests) classified by the interest areas. It can be created by any user, what matters is a choice of rubric (sports, news, gaming, etc.). Subreddits are moderated, that is why some discussion sections of the rubric may fail due to the inappropriate statements in the comment sections. There is a possibility for each user to provide and fill in a content, to post and comment or submit a comment to another comment depending on personal attitude and preferences [11, 12].

The Reddit social network implies an attempt of modeling and gaining insights into online conversation patterns. Recent studies concerning Reddit include such interest areas as behavioral attitudes towards community users [13], models providing boosting processes in terms of comment sections of the online platform [14, 15], as well as bridging the gap in terms of content relevancy [10]. It has been also discovered that the content emerging at its latest seems to be more attracting and commented [9]. It has been shown that the posts made for the first time are also commented more often than those which appeared later [15]. These cases refer to the psycholinguistic features, and can be interpreted in the following way: in any situation when it comes to make a right choice, there is a kind of “rule” which triggers only when totally new information (verbal, visual, sensual, perceptual, etc.) emerges and at that same time our brain reacts, it is ready to act. The simple reason – we always seek “fresh” unexpected findings. The same situation concerns posts in social networks: there is a high probability that those people who have just joined the online community will implicitly comment the latest post and their attention will be attached to the newest image from the top without scrolling down to the bottom of the page.

Psycholinguistic studies are related to the relationship between human mind and language [16]. This branch of linguistics comprises cognitive processes of language studies: “language as a paradigmatic system, that is, a set of choices for each instance

from which a speaker must select one. Such a set of choices is inherently probabilistic, that is to say in each situation, various choices are more or less likely to be selected by a speaker” [17]. When an author of the online community refers to a certain cognitive process explicated in the statements, there is an intention to state expressively the issue arisen, to attract other users to the commenting system of the network.

Nowadays, the most widespread reason why people cause conflict situation during conversation is fruitless dialogues or polylogues having no purpose of speech and which are considered to be unnecessary. The notion of conflict, conflict situation and contradictions are investigated in terms of contact linguistics [18], sociolinguistics [19], psycholinguistics [20].

3 Revealing Psychological and Linguistic Features of the Latest Reddit News Post

When there is a climax of a conflict situation, communicators have to find all possible solutions to the problem. The most requiring is to be approachable and flexible while communicating – you will be definitely able to support the relevant conversation atmosphere and avoid sharp opinions, managing the conflict situation. Moreover, freedom of speech is welcome any time – interlocutors should encourage each other and care about tone of conversation supporting contradictory statements with mitigating phrases and positive emotions. If the contradictions are inevitable way within conversation, it is better to discuss the issue with an attempt to solve it immediately. Maybe, while discussing, people agree on some other important points and will forget about the previous contradictory issues. Anyway, it is advisory to bear in mind that any beginning of conversation has its positive and negative outcome.

As it was mentioned in previous papers, the best model of dealing with any conflict situations, contradictory statements is the following chart of crucial conflict resolution strategies (<https://kilmanndiagnostics.com/overview-thomas-kilmann-conflict-modeinstrument-tki/>) (Fig. 1). Notwithstanding the fact that they are dominantly in oral speech, we apply them to the textual space in terms of online communities.



Fig. 1. Crucial conflict resolution strategies

Accordingly, the aim of this paper is to verbally reveal cognitive processes with their negative and positive implications using linguistic markers within the selected discussion comment sections in the Reddit news post and compare them applying a text analysis

program detecting linguistic markers and psychological patterns – Linguistic Inquiry and Word Count (LIWC, <https://liwc.wpengine.com/results/>). Additionally, there is a task to compare the analysis of both previous papers [11, 21] with this one concerning conflict nature of the comments of Reddit posts in terms of linguistic interpretation regarding crucial conflict resolution strategies. What is more, here we apply the conflict resolution strategies to the news posts and their comments. Also, we apply LIWC analysis, both previously manual and automatic, to the early paper [21]. All the findings are to be linguistically analyzed and compared.

LIWC can detect more than 2,000 words classifying them into linguistic, psychological, social and personal attitudes [22]. The report of the program represents the percentage of the words totally enclosed in the chart for thorough analysis [23]. The results of the automatic analysis provided by LIWC do not claim 100% true results, though it depends on the sustainable amount of words (more than 500) for the sample and, as a result, on the verbal frequency, which can be counted manually. In this case, such an attempt may be time-consuming.

As it has been stated above, Reddit social platform is chosen for the linguistic textual analysis of its content. Random news rubric of the Reddit has been allocated. It concerns the death of Irfan Khan, famous Indian actor (https://www.reddit.com/r/movies/duplicates/ga4cve/irrfan_khan_actor_extraordinaire_and_indias_face/).

The reason to choose such news was the psycholinguistic trigger process described above – the first news the visual channel perceived and processed. The rubric section was automatically marked as news post (Fig. 2). The first part of comment section following the rubric section with the next comments (Fig. 3). As the first comment section has been structurally described and exemplified before [11], here, an accent is made on discussion section with the following comments (Fig. 4).

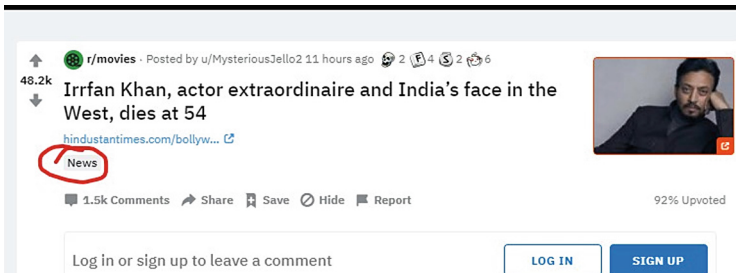


Fig. 2. The Reddit news post

It is represented by 12 discussion charts; each has its own comments. The psycholinguistic trigger makes sense here as well – the first six discussion charts attract users' attention in terms of different titles of the charts (the effect of "What if I find missing information" plays a crucial role when scrolling down the charts of the discussion section). Of course, the news post is verbally marked as an expression of grief, but still the alteration in its formulating requires different personal attitudes. Each comment statement is also marked with an appropriate resolution strategy mode, which will be analyzed in further paragraph of this paper.



Fig. 3. The extended comment section followed by the next comments



Fig. 4. Discussion section with the following comment

The first discussion chart in the following discussion section to consider is marked as non-political section “Irrfan Khan dies at 54” (Fig. 5) with the following linguistic layout markers (the following semantic groups are allocated in accordance with those introduced by LIWC) expressing:

12 OTHER DISCUSSIONS

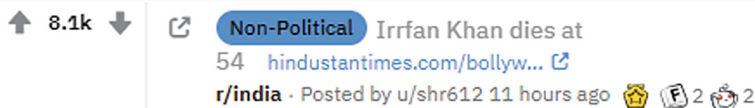


Fig. 5. The first discussion chart marked as non-political section “Irrfan Khan dies at 54”

- **negative emotions:** *terribly sad news* (grief), *Sad news* (grief), *condolences* (grief);
- **positive emotions:** *one of the most talented actors* (admiration), (collaborating), *What a man* (collaborating), *what an actor* (admiration), (collaborating), *Never heard anything negative about him* (admiration), (collaborating), *he's the best actor this country has ever produced* (admiration), (collaborating), *His confidence on the screen was unparalleled* (admiration), (collaborating)...;
- **cognitive processes:** *Rest in Peace Irrfan Khan* (implicitness of death), *Gone too soon* (implicitness of death), *Just gone too soon* (implicitness of death), *we shall meet* (implicitness of death) *again* (implicitness of death), *Irfaan Khan salute* (implicitness of death);
- **reference to personal factor:** *I remember* (past reference). Such textual allocation and semantic “customizing” is really time-consuming, though it must be stated that cognitive processes interrelate with emotional sphere, that is why manually it is complicated to attribute linguistic markers to the 100% exact allocation. However, statistically, the program representation, it has the same result. After thorough textual analysis of the linguistic markers we prove the manual result with an automatic report (Fig. 6).



TRADITIONAL LIWC DIMENSION	YOUR DATA	AVERAGE FOR SOCIAL MEDIA: TWITTER, FACEBOOK, BLOG
I-WORDS (I, ME, MY)	2.2	5.51
SOCIAL WORDS	14.8	9.71
POSITIVE EMOTIONS	5.1	4.57
NEGATIVE EMOTIONS	3.6	2.10
COGNITIVE PROCESSES	12.3	10.77

Fig. 6. Percentage overview of the LIWC automatic analysis of the semantics of the Reddit comments in the first discussion chart

Comparing the average percentage for social media, Reddit’s comments to the news post represent more frequent usage of positive emotions than in general social platforms, and the same concerns negative emotions. Such exaggeration in textual comments prove the admiration for the actor which is equal to his early death. Further percentage reports, presented in Fig. 5 will be automatically provided in Table 1 below.

It is essential to note that the first discussion comment chart is the most contentdriven, as it was “attacked” by users immediately, and all the rest charts are full of repetitive comments in different interpretations.

The second discussion chart in the following discussion section to consider is marked as India-speaks news section “Irrfan Khan, actor extraordinaire and India’s face in the West, dies at 54 - bollywood” with the following linguistic layout markers expressing:

Table 1. Table of frequency of LWIC automatic textual analysis of the Reddit comments in the discussion charts

Comment section	I-words (I, me, my)	Positive emotions	Negative emotions	Cognitive processes
1	2.2	5.1	3.6	12.3
2	2.6	6.5	4.7	10.5
3	2.6	10.5	10.5	18.4
4	0.0	5.3	5.3	13.2
5	2.1	4.2	6.3	4.2
6	1.6	7.9	11.1	11.1
7	0.0	20.0	20.0	10.0
8	6.7	0.0	3.3	10.0
9	0.0	6.3	6.3	0.0
10	0.0	0.0	0.0	21.4

- **negative emotions:** *the first time I am feeling sad about Bollywood* (grief + **personal factor**), (collaborating);
- **positive emotions:** *Bollywood doesn't have much talent. Probably can count with just the fingers of my hands. Irrfan was **leading that bunch*** (admiration), (collaborating), *He was a **gem of an actor*** (admiration), (collaborating), ***what a wonderful actor*** (admiration), (collaborating);
- **cognitive processes:** *May their souls rest in peace* (implicitness of death), *the **other Khans are worthless and talentless*** (implication of admiration), (competing), *I am sure they meet in heaven* (+**personal factor**);
- **reference to personal factor:** I hope (reference to the future). Here, we prove with an automatic program report that there are more positive emotions in the comments than negative ones.

The third discussion chart in the selected discussion section to regard is marked as movies news section “Indian actor Irrfan Khan (Jurassic World, Amazing Spider Man, Slumdog Millionaire) has died” with the following linguistic layout markers expressing:

- **negative emotions:** *I didnt not expect that, pretty sad* (grief + **personal factor**), (avoiding);
- **positive emotions:** *Damn, **loved him** in so many movies* (admiration), (collaborating), *A **genuinely talented and universally admired actor*** (admiration), (collaborating);
- **cognitive processes:** ***Shit...didn't know** he was sick* (implicitness of death + grief), (avoiding). In this allocation, we prove with an automatic program report that there is a relevant percentage between positive and negative emotions in the comments.

The fourth discussion chart in the following discussion section to focus is marked as “Indian actor Irrfan Khan, from The Amazing Spiderman, Life of Pi dies at 54 due to colon infection” with the following linguistic layout markers expressing:

- **positive emotions:** *He is a big deal here in India* (admiration), (collaborating), *Very*
- **versatile actor and a genuinely good person** (admiration), (collaborating);
- **cognitive processes:** *What the FUCK 2020!!* (implicitness of desperate condition), (avoiding), *but this one hurt* (explicitness of grief), *may he rest in peace* ♥ (implicitness of death). In this group, we prove with an automatic program report that there are deep implicit cognitive processes in the comments.

The fifth discussion chart in the chosen discussion section to emphasize is marked as Bollywood news section “Irrfan Khan dies at 54” with the following linguistic layout markers expressing:

- **negative emotions:** *This is genuinely sad* (grief), (collaborating);
- **positive emotions:** *he was at the height of his career* (admiration), (avoiding), *Amazing actor Legendary filmography* (admiration), (collaborating), *unironically one of the better human beings and actors in a shit filled industry* (admiration), (collaborating and competing);
- **cognitive processes:** *2020 is officially cancelled* (implicitness of desperate condition), (avoiding), *This has genuinely ruined my day* (implicitness of desperate condition), (avoiding). In such allocations, we have the same title but contrastive comments
- manually, cognitive processed are implicitly more deeply expressed than the program represents.

The sixth discussion chart in the allocated discussion section to pay attention to is marked as film general news section “Irrfan Khan, actor extraordinaire and India’s face in the West, dies at 54” with the following linguistic layout markers expressing:

- **negative emotions:** *It’s unfortunate that he had to die after battling so much pain and suffering* (grief), (collaborating);
- **positive emotions:** *A grave loss to the entire world of acting* (admiration), (avoiding), *He was one if those guys I didn’t see much in the US but always made whatever character he was playing interesting* (admiration), (avoiding, collaborating);
- **cognitive processes:** *The world lost one of its best actors today* (implicitness of death), (avoiding, collaborating), *RIP* (implicitness of death). As we can see, we have explicitly profound linguistic negative markers, which is proved in percentage in LIWC.

The seventh discussion chart in the stated discussion section to focus is marked as news section “Indian actor Irrfan Khan dies at 54” with the following linguistic layout markers expressing:

- **negative emotions:** *Terribly sad* (grief);

- **positive emotions:** *He was a truly gifted actor and entertainer* (admiration), (collaborating). It is a short comment section but true in negative and positive emotions.

The eighth discussion chart in the allocated discussion section to claim is marked as Kerala news section “A big loss.. RIP” with the following linguistic layout markers expressing:

- **negative emotions:** *Can’t fathom he is gone* (grief), (avoiding);
- **cognitive process:** *I don’t know how I’ll be able to walk into a theater and see his face again* (implicitness of desperate condition), (avoiding); *This year is a nightmare* (implicitness of desperate condition), (avoiding), *RIP Irrfan Khan* (implicitness of death). It is a short version of comment section but true in implicitly profound cognitive processes negative emotions.

The ninth discussion chart in the below discussion section to emphasize is marked as wesanderson news section “Irrfan Khan (“The Father” from Darjeeling Ltd), actor extraordinaire and India’s face in the West, dies at 54” with the following linguistic layout markers expressing:

- **negative emotions:** *We lost him that way too soon* (grief), (avoiding), *Oh no* (grief), (avoiding);
- **positive emotions:** *One of the generation’s finest Indian performers* (admiration), (collaborating). It is too short variant of comment section but it has the same percentage measurement in positive and negative emotions.

The tenth discussion chart, the last one in the discussion section, is marked as howardstern news section “Hanzi has Passed Away” with the following linguistic layout markers expressing cognitive clarification: *Imran Khan is his name Not Irrfan* (competing), which does not refer mainly to the topic discussed.

The program proves manual textual analysis – none of the allocations are mentioned.

All in all, each of ten discussion comment section charts is marked with linguistic implications and explications as well. The first six comment sections proved to be the most commented, the rest four – the least. Each has its own linguistic content, where mainly dominated negative emotions as it is a grief and positive emotions as it was a great actor. Cognitive processes involved personal factor and are marked with deep implicit structure. In the fifth comment, to our mind, manual textual analysis of linguistic markers is more exact manually than when it was analyzed using the program. So, the percentage results prove actor’s profound fame and unexpected death (see Table 1).

4 Analysis of the Reddit Conflict Situations

Both previous papers concerning the conflict nature of comments of the Reddit social network posts have been thoroughly analyzed in terms of textual analysis using linguistic markers according to crucial conflict resolution strategies (Table 2, 3).

In order to provide comparison in terms of three recent news comments of the Reddit posts, let us analyze the comment section of this paper’s news about the death of Indian

Table 2. Table of frequency of conflict resolution strategies of the Reddit news conversations about minorities and their right to obtain scholarship.

Comment section	Accommodating	Collaborating	Avoiding	Compromising	Competing
1	3	6	4	1	3
2	0	0	0	1	0
3	0	0	1	0	3
4	0	1	1	0	0
5	1	2	0	0	0

Table 3. Table of frequency of response conflict modes of the Reddit news conversations about Coronavirus.

Comment section	Compromising	Accommodating	Competing	Avoiding	Collaborating
1	0	1	2	2	2
2	0	2	3	1	1
3	0	1	1	1	4
4	0	2	3	0	2

actor. The comments were marked with an appropriate strategy mode according to the statements of the comments in the previous paragraph of this paper. However, we should accept the fact that news about the death cannot be controversial in its nature, still, we consider it in terms of a conflict situation as there is a polylogue within the Reddit discussion section and communicators reply with different mode of assertiveness and emotions triggering opposite reactions of other users, thus, causing contradictions, even concerning such a grief news.

What has to be noted is that mainly techniques of collaborating have been applied to the statements of the communicators, as mostly each and every one supported mourning emotions within the discussion, which was expressed and defined by such linguistic markers of desire to share grief, support to overcome sad situation – contradictory one in terms of his early age and sudden death. When interlocutors sounded contradictory, competing technique of resolution strategy mode was applied to the comments, as they reply explicitly not collaborating with the previous statements but introducing their own opinion. Avoiding technique was used in those comment statements when communicators’ reactions were totally negative not praising or remembering actor’s lifetime. So, the Table 4 is the following.

When comparing three tables of frequency of response conflict modes of the Reddit recent news conversations (the first is about giving scholarships based on minority status, the second – Coronavirus, the third – the death of Irfan Khan), we must note that the first two news refer to the conflict situation notion, but the last one is controversial in its nature, not by its conversation model. Nonetheless, we mark the latter as news

Table 4. Table of frequency of response conflict modes of the Reddit news conversations about the death of Indian actor.

Comment section	Compromising	Accommodating	Competing	Avoiding	Collaborating
1	0	0	0	0	6
2	0	0	1	0	4
3	0	0	0	2	2
4	0	0	0	1	1
5	0	0	1	3	3
6	0	0	0	3	3
7	0	0	0	0	1
8	0	0	0	3	0
9	0	0	0	2	1
10	0	0	1	0	0

post, it has a polylogue structure, people reply to each other expressing their approval and indignation, trying to show their best in linguistic way of paraphrasing mourning statements, even advocating their rightness. In comparison with two other Tables (9 and 9 collaborating modes in both news discussion sections), the third one has the biggest amount of comment Sects. (10). Surprisingly, we define the last one as the least contradictory by its conversation section, still, it has the biggest number of collaborating techniques (21) comparing to the other news discussion sections. The same situation we have with an avoiding technique – the most cases of application (14) of this mode we have in the third news discussion section in comparison with 6 and 4 cases of the previous two cases.

The application of competing technique to the first two news discussion section gave unexpected results: Table 2 (6) has less cases of this mode than Table 3 (9). It is linguistically and statistically proved as the relevant topic of Coronavirus is controversial indeed. There were a lot of conflict “attacks” which were defined by linguistic markers of negative contradictory tones.

The usage of collaborating techniques is the same in amount (9 and 9 cases), as the manner of communicators were pretty the same in terms of trying not to stand out within the conversation, to show respect and mind each other’s comments.

The comments marked with the avoiding techniques are relevantly similar (6 and 4 cases) as users of both news discussion sections in terms of conflict situation tend to leave contradictions unresolved, try to end the conversation with a rhetorical question or pretend to unexpectedly paraphrase the previous statement and disengage.

The compromising technique was rarely applied to the mentioned news discussion sections (only 2 cases in the first one). It is substantiated that online news discussion usually cannot be compromised in any conflict situation. It ends up with unresolved issues, sharp arguments, rhetorical questions or unwillingness to continue the conversation.

Accommodating technique is characterized by 4 and 6 cases of its application in terms of news discussion sections, what cannot be said about Table 4 (0 cases), as no one can adapt to grief news. The first news about the attitude towards scholarship in minorities was considered to be more “positive” comparing with more “negative” second Coronavirus news. It means that it was easier for some users to adapt, i.e., to agree with the statements without introducing their own opinion.

The most distinctive feature of the third Reddit news comment section that we discovered is that some statements tend to be divided into two parts. The first part is defined by one resolution strategy technique, and the other one – by another one. It is an evidence of the fact that at first, communicators tend to, e.g., approve of the previous statement, and continue to express his/her own opinion, at the beginning of the statement, interlocutor prefers to abruptly change the way of thinking and then tries to mitigate the conflict situation and collaborate.

5 Conducting the Semantic Analysis of Both Previous Reddit News Posts and Their Comparison

In this paragraph, we apply textual semantic analysis in order to allocate the following semantic groups in both comment sections of the previous papers and compare them with the one analyzed here in the third paragraph of this paper. The first discussion chart in the following discussion section to consider is marked as “Giving scholarships based on minority status” (https://www.reddit.com/r/unpopularopinion/comments/dbtzpg/giving_scholarships_based_on_minority_status/) with the following linguistic layout markers expressing:

- **reference to personal factor in the first comment section:** *I’m not even sure* (explicitness of personal doubts), *I thought the same thing too* (explicitness of repetitiveness), *I think we all think this is interesting* (explicitness of team spirit), *I see your point* (explicitness of understanding);
- **reference to personal factor in the second comment section:** -
- **reference to personal factor in the third comment section:** *As far as I’m aware* (explicitness of personal opinion);
- **reference to personal factor in the fourth comment section:** -
- **reference to personal factor in the fifth comment section:** -
- **negative emotions in the first comment section:** *I’m not even sure how this affects the conversation* (implicitness of bad effect), *It’s not very relevant* (explicitness of inappropriateness), *the gap still will continue* (explicitness of inability);
- **negative emotions in the second comment section:** -
- **negative emotions in the third comment section:** *You know what’s also bullshit? ..more offensive than the other?* (explicitness of rudeness), *.. getting free college because they’re rich?* (implicitness of indignation), *that really didn’t answer my question* (explicitness of inability);
- **negative emotions in the fourth comment section:** *The sad thing is that* (explicitness of misery);
- **negative emotions in the fifth comment section:** -

- **positive emotions in the first comment section:** *That would actually solve the problem* (explicitness of hope), *to help bridge the gap* (explicitness of hope), *Not gonna lie* (explicitness of true nature);
- **positive emotions in the second comment section:** -
- **positive emotions in the third comment section:** *More importantly how* (explicitness of intensity);
- **positive emotions in the fourth comment section:** -
- **positive emotions in the fifth comment section:** *thanks for the interesting perspective* (explicitness of gratitude);
- **cognitive processes in the first comment section:** *Some people feel that way, others don't* (explicitness of obvious life situation), *Something to consider* (explicitness of contemplation), *Nobody cares what they think* (explicitness of obvious life situation), *believe it or not* (explicitness of attitude), a **huge assumption** (implicitness of doubtful occurrence), *This was **certainly** a fact!* (explicitness of assertive opinion), *You've got to be **rich** to even **think** that way* (explicitness of condition);
- **cognitive processes in the second comment section:** -
- **cognitive processes in the third comment section:** -
- **cognitive processes in the fourth comment section:** -
- **cognitive processes in the fifth comment section:** *I completely agree with you* (explicitness of approval).

After conducting such a detailed analysis, we can state that in the first comment section of the mentioned rubric: the biggest amount of cases (7) refer to the semantic group of cognitive processes, as the topic of the news discussion is really requiring and the issue about limiting scholarship in minorities is really of importance. The semantic markers of common life situations prevail. Obvious, explicit ways of representing information dominate. The semantic group of reference to personal factor (4 cases) comes after cognitive processes involving common personal reactions. Negative and positive emotions are statistically the same (3 cases). The same results are revealed applying the automatic program (Table 5).

Table 5. Table of frequency of LWIC automatic textual analysis of the Reddit comments in the discussion charts

Comment section	I-words (I, me, my)	Positive emotions	Negative emotions	Cognitive processes
1	3.0	5.9	3.6	23.7
2	–	–	–	–
3	5.4	8.1	2.7	29.7
4	7.1	0.0	7.1	0.0
5	4.8	14.3	0.0	14.3

In the second comment section, we do not apply the automatic program, as the statements were not allocated and referred to any of the semantic groups.

In the third comment section, the results of applying automatic analysis are surprising, as in some semantic groups we allocated only one or none of the statements, but the program counted some other implicit cognitive processes concerning total emotions within the conversation. In this case, manual analysis is more accurate.

In the fourth section, manual and automatic analysis coincide with their results. The only semantic group – reference to personal factor – can be misleading, as the program takes into account all personal pronouns available within the statements.

The last comment section also proves similar results of both manual and automatic analysis. But the same situation as in the previous comment section is with the semantic group of reference to personal factor.

The second discussion chart in the following discussion section to consider is marked as “Coronavirus Megathread” (https://www.reddit.com/r/unpopularopinion/comments/fhrs6r/coronavirus_megathread/) with the following linguistic layout markers expressing:

- reference to personal factor in the first comment section: -
- reference to personal factor in the second comment section: -
- reference to personal factor in the third comment section: -
- reference to personal factor in the fourth comment section: -
- negative emotions in the first comment section: *have not been giving consistent messaging* (explicitness of inappropriateness), ...*I can't* in good conscience *vote red next election* (explicitness of inability), *I just can't* (explicitness of inability), *They both fucked up* (explicitness of rudeness), *not to use the WHO test* (explicitness of inability), *ignorant and incompetent* (explicitness of rudeness), *Americans need to pull Trump out of the White House tonight* (explicitness of rudeness), *That's a scary opinion since it would violate our democracy* (explicitness of inappropriateness),
- negative emotions in the second comment section: *All this talk of “selfishness” and “hysteria” is just their way of coping with the anger of being late to the party* (implicitness of misery), *It is not my problem you waited* (implicitness of rudeness), *It's selfish* (explicitness of misery), *That's heartless* (explicitness of misery), *These masks are NOT medical masks and will do NOTHING to stop a virus ..So now work that needed to be done isn't being done* (explicitness of inability and inappropriateness), *That seems like a supply chain problem* (implicitness of inability),
- negative emotions in the third comment section: *I'm not glad that people are dying and suffering in any way because of COVID-19* (explicitness of misery), *Coronavirus is a good thing because all the healthy people are coming out alright or better, and all the old people and people in poor health who can't contribute to the economy or the better advancement of humanity are dying off* (explicitness of common life rudeness), *Why should I care if some 80-year-old who has lived their life, but continues to live because modern medicine refuses to let them die, dies off?* (explicitness of common life rudeness), *Is it sad? Not really. A lot of people die every day* (explicitness of common life rudeness), *Do I like him? No* (explicitness of common life rudeness),
- negative emotions in the fourth comment section: *This is stupid on so many levels* (explicitness of rudeness), *Everyone should get infected by the coronavirus? Are you*

willing to **submit yourself to that**? (implicitness of internal misery), .. *Or maybe I should live as watch the ones u love **die one by one. Yeah second choice*** (implicitness of rudeness)

- **positive emotions in the first comment section:** *the most **coherent, fair and informative*** (explicitness of trustworthiness), ***rare breed*** (implicitness of trustworthiness),
- **positive emotions in the second comment section:** ***Stop whining, be proactive** about being prepared for emergencies* (explicitness of cheerfulness), *Only **anticipate the actual need and stretch it as much as possible. Don't over stock** and say, "Well, it's your fault **I made it there before you*** (explicitness of cheerfulness), *Thats **awesome**..* (explicitness of cheerfulness),
- **positive emotions in the third comment section:** *But to be **optimistic** our country need to **not give up and not think things will be hard*** (explicitness of cheerfulness)
- **positive emotions in the fourth comment section:** *Everyone who lives **develops immunity** and the **virus** is completely **wiped out*** (explicitness of cheerfulness), *Sounds **good** to me* (explicitness of cheerfulness),
- **cognitive processes in the first comment section:** *Why are you **blaming Trump** when you just laid out exactly* (explicitness of wonder), *since he's the **boss, he is accountable too*** (implicitness of attitude),
- **cognitive processes in the second comment section:** ***Healthy individuals** are*
- ***wayyy overpurchasing*** (explicitness of attitude),
- **cognitive processes in the third comment section:** *The truth is this **thing is serious*** (explicitness of attitude), *there are **truths** to what people are saying **about this virus and disease*** (explicitness of attitude), *Again this **isn't war** but this situation **isn't a joke*** (explicitness of attitude)
- **cognitive processes in the fourth comment section:** *Those who **survive, survive** and those who **die, die*** (explicitness of attitude), *..**being unpopular** only **affects** yourself.* (explicitness of attitude), *Yeah **sign me up!** Where's the **closest** person you know with **Coronavirus?** I'll **square up*** (explicitness of attitude)

After thorough semantic analysis conducted above, we can assume that in the first comment section of the mentioned rubric: the semantic group of negative emotions has the biggest amount of cases (8) proving the actual state of the world's conditions in terms of pandemic. The semantic groups of positive emotions (2 cases) and cognitive processes (2 cases) come after the group of negative emotions. The program analysis proves the manual results, although it identified personal pronouns, as in previous research, and cognitive processes are calculated in general including negative and positive emotions. So, manual allocation is more accurate (Table 6).

In the second comment section, the semantic group of negative emotions has the biggest number of cases (6), the results of manual analysis prove those of automatic program, except for the first semantic group – reference to personal factor.

In the third comment section, the results of applying automatic analysis are surprising, as still negative attitudes and emotions prevail. The program may take into account all words that occur within the negative implicit nature of the statement. In this case, manual analysis is more accurate.

In the last comment section, manual and automatic analyses do not coincide with their results. The semantic group of negative emotions – 5 cases, positive - 2 cases. But

Table 6. Table of frequency of LWIC automatic textual analysis of the Reddit comments in the discussion charts

Comment section	I-words (I, me, my)	Positive emotions	Negative emotions	Cognitive processes
1	3.3	4.7	6.0	19.3
2	1.7	2.3	5.1	10.2
3	3.4	7.8	3.4	18.1
4	3.1	2.5	1.3	11.9

manually a group with negative emotions is semantically analyzed and considered to be more accurate than the automatic analysis. Moreover, semantic group – reference to personal factor – can be misleading, as the program takes into account all personal pronouns available within the statements.

Finally, let us compare the results of frequency of LWIC automatic textual analysis of the Reddit comments in the discussion charts (total amount is provided in Table 7).

Table 7. Table of total frequency of LWIC automatic textual analysis of the Reddit comments in the discussion charts

News discussion charts	I-words (I, me, my)	Positive emotions	Negative emotions	Cognitive processes
1	20.3	28.3	13,4	67.7
2	11.5	17.3	15.8	59.5
3	17.8	65,8	71.1	111.1

Firstly, we should note that research is approved when the results of manual and automatic textual analyses coincide. However, such a profound and detailed semantic analysis allowed us to state that manual results tend to be more precise, as important linguistical nuances were taken into consideration while interpreting the textual space within the Reddit social network. In such cases where LWIC showed more or less precise results, we can be sure that manually we proved it using linguistic markers and interpretative approaches.

So, the first semantic group that was allocated is reference to personal factor. It is considered to be the most misleading in manual analysis, as we reveal implicit nature of the statements, and some personal pronouns were not referred to that group, although we may be sure that LWIC counted them appropriately. As we can see, the total frequency of LWIC automatic textual analysis of the Reddit comments in the discussion charts was provided in final Table 7. Of course, it is unfair to interpret results in total, as there were different numbers of comment sections, still, in brief, we can apprehend a general overview. The amount of general cognitive processes in the third news discussion chart is substantiated by its number of comment sections. It is noteworthy to say that

more amount of cognitive processes prevailed in the first Reddit news as it was more intellectual topic to discuss than coronavirus emergency.

The total amount of negative emotions is precise as manually as automatically: Coronavirus pandemic is totally negative situation nowadays, but mourning news about sudden death is dramatic.

Positive emotions illustrate the biggest amount of comments devoted to prolific actor in the third Reddit news discussion chart, and hopes for possibility to gain scholarship in minorities in the second item of news discussion chart.

6 Conclusion

In general, elements of linguistic, psychological, textual semantics analysis were closely interconnected in the paper. Social network platforms were claimed to be a medium of immediate information sharing, commenting and posting processes. The Reddit social community with its extended comment structure was allocated. The Reddit comments were marked linguistically, grouped, analyzed and proved statistically using the automatic LWIC program for detecting linguistic markers and psychological patterns of human being. The linguistic comments followed the allocated groups, and, finally, the table of frequency of response conflict modes of conversations enabling to apprehend the comments' deep structure statistically was provided.

As it was an attempt to compare results of both previous papers concerning the notion of conflict situation and contradictions within the social media, the aim of thorough textual semantic analysis of previous textual material was achieved. Moreover, the conflict resolution strategies were applied to the news posts and those comments, which were previously allocated. Furthermore, the automatic LIWC analysis of both early Reddit news comment sections was conducted.

In perspective, the extended version of Reddit comment system in different areas may be considered and used as a basis for further studies in the field of global social networking.

References

1. International Communication Union. ICT Facts and Figures (2016)
2. MicroFocus, How Much Data is Created on the Internet Each Day [Internet], 8 September 2016
3. Internet Live Stats. Twitter Usage Statistics [Internet]
4. Benevenuto, F., Rodrigues, T., Magno F., Almeida, V.: Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, pp. 49–62 (2009)
5. Gómez, V., Kaltenbrunner, A., López, V.: Statistical analysis of the social network and discussion threads in Slashdot. WWW (2008)
6. Kumar, R., Mahdian, M., McGlohon, M.: Dynamics of conversations. ACM KDD (2010)
7. Marcoccia, M.: On-line polylogues: conversation structure and participation framework in internet newsgroups. *J. Pragmatics* **36**(1), 115–145 (2004)
8. Mayfield, E., Adamson, D., Rosé C. P.: Hierarchical conversation structure prediction in multi-party chat. In: The 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (2012)

9. Gilbert, E.: Widespread underprovision on reddit. In: ACM CSCW (2013)
10. Singer, P., Flock, F., Meinhart, C., Zeitfogel, E., Strohmaier, M.: Evolution of reddit. From the front page of the internet to a self-referential community. In: WWW Companion (2014)
11. Albota, S., Peleshchyshyn A.: Contradictory statement as a basis for conflict resolution strategies. In: CEUR Workshop Proceedings, Conflict Management in Global Information Networks. CMiGIN 2019, vol. 2588, pp. 336–345 (2020)
12. Khomytska, I., Teslyuk, V., Holovatyy, A., Morushko, O.: Development of methods, models, and means for the author attribution of a text. *Eastern-Eur. J. Enterpr. Technol.* **3/2**(93), 41–46 (2018)
13. Das, S., Lavoie, A.: The effects of feedback on human behavior in social media. An inverse reinforcement learning model. In: The International Conference on Autonomous Agents and Multi-agent Systems (2014)
14. De Choudhury, M., De, S.: Mental Health Discourse on reddit: self-disclosure, Social Support, and Anonymity. In: ICWSM (2014)
15. Weninger, T.: An exploration of submissions and discussions in social news: mining collective intelligence of reddit. *Soc. Network Anal. Min.* **4**(1) (2014)
16. Maftoon, P., Shakouri, N.: Psycholinguistic approach to second language acquisition. *Int. J. Lang. Learn. Appl. Linguist. World (IJLLALW)* **1**(1), 1–9 (2012)
17. McEnery, T., Wilson, A.: *Corpus Linguistics: An Introduction*. Edinburgh University Press, Edinburgh (2001)
18. Brekhus, W. H., Ignatow, G.: *The Oxford Handbook of Cognitive Sociology* (2019)
19. Ahmed, S., Cho, J., Jaidka, K.: Framing social conflicts in news coverage and social media. A multicountry comparative study. *Int. Commun. Gazette* **81**(4), 346–372 (2019)
20. Abdel-Fadil, M.: The politics of affect. The glue of religious and identity conflicts in social media. *J. Religion, Media Dig. Cult.* **8**(1), 11–34 (2019)
21. Albota, S.: Resolving Conflict situations in reddit community driven discussion platform. In: Computational Linguistics and Intelligent Systems. COLINS. CEUR Workshop Proceedings, vol. 2604, pp. 215–226 (2020)
22. Pennebaker, J., Francis, M. E., Booth, R.J.: *Linguistic inquiry and word count (LIWC). A computerized text analysis program*, Mahwah, NJ, Erlbaum (2001)
23. Pennebaker, J., Mehl, M. R., Niederhoffer, K. G.: Psychological aspects of natural language use. Our words, our selves. *Ann. Rev. Psychol.* **54**, 547 (2003)



A Model of the Information System of the Associative Verbal Network Presentation

Olena Levchenko¹ , Oleh Tyshchenko¹ , Marianna Dilai¹ ,
and Lukas Gajarsky²

¹ Lviv Polytechnic National University, Lviv 79013, Ukraine
levchenko.olena@gmail.com, olkotyszczenko@gmail.com,
mariannadilai@gmail.com

² University of Ss. Cyril and Methodius in Trnava, 917 01 Trnava, Slovakia
lukasajarsky@ucm.sk

Abstract. This study presents a method of modeling the associative verbal network (hereinafter AVN) on the basis of the associative test data and, accordingly, reveals the principles of building an information system for presenting AVN of native speakers of different languages from typological and cognitive perspectives.

This technique involves a comparative analysis of AVNs of the studied languages: identifying the specifics of associative content exemplified by the Ukrainian and Slovak conceptual domains ЗРАДА/ZRADA ‘BETRAYAL’, their taxonomy and typology of related frames, subframes and interconceptual relations (areas of semantic intersection, associative similarity, degree of remoteness of verbal responses in nuclear and peripheral areas of the associative field, formalization and visualization by lingual cognitive, statistical procedures and methods of automatic information processing with the involvement of related interdisciplinary areas (sentiment analysis, frame modeling, semantic profiling, etc.).

A key component of this technique is the ‘associative’ distance between concepts determined by analyzing data on common associations (index of mutual associative relation), as well as visualizing the associative test data, which allows us to identify such common areas.

It is concluded that this approach opens new typological opportunities and prospects for building a comparative model of the associative network, reflecting the facts of the Slavic mentality and their possible modeling on cognitive and axiological, quantitative-parametric basis.

The paper describes the modules of the information system for AVN presentation.

Keywords: Information system · Associative verbal network · Index of mutual associative relation · Associative test · Semantic distance

1 Introduction

In modern linguistics, in particular psycholinguistics, cognitive linguistics, intense discussions are held regarding research methods [1] and methods of formalization, cognitive and computational processing, explanation of empirical linguistic facts, including culturally marked ones.

In contrast, N. Vasilieva believes that “when conducting an associative test, frames /schemes ... situations as the main forms of representation of knowledge are activated” [2]. Most likely, a person, being in a situation of associative experiment, is focused on the mode of “communication with the environment”, but not on the mode “for the self” [3], which is more common for mental speech, although it is manifested in scarce reactions.

Modern technological advances in building semantic networks such as WordNet [4] and GlobalWordNet [5] “do not completely replace associative dictionaries and databases, but move the data obtained as a result of the associative test to the periphery of the study of language ability and speech activity. Associative dictionaries remain a source of studying the individual variability of communicative competence and cognitive development” [6].

Thus, the associative test provides non-trivial data that are significant not only for the linguistic studies per se. It should be noted that the data about the interconceptual associative distance between stimulus words, in our opinion, can be used to verify and refine the results of the so-called distributive models aimed at identifying the semantic distance between words, which in turn are used in word clustering based on their semantics; to generate thesauri and multilingual dictionaries; to expand queries in information retrieval; to determine the theme of documents; to build semantic maps of certain subject areas [7].

The associative test results are used in different fields [8–15]. A number of attempts to visualize certain fragments of associative tests have been presented [16]. Unfortunately, the results of associative experiments are presented in digital form only for some languages (XML version of Edinburgh Associative Thesaurus (EAT) [17], Russian Associative Dictionary [18], etc.).

However, they do not present visualization, taxonomy and contrastive analysis of responses in different languages. Another project is called Word Associations Network (does not support the Ukrainian language), but it is not based on the associations obtained experimentally from respondents: “The heart of the process of forming the list of associations is a software module that analyzes the classical and contemporary works of English literature using the key principles of systems approach. The combination of unique algorithms developed by the author of the project, lets you to contemplate a set of associations with a given word. The algorithm design took into account the specific processes in the human brain. Therefore, the created list of associations can be considered as the average result of the implicit association test» [19]. The authors claim that Word Associations Network is inherently an ideographic dictionary or thesaurus [19].

2 Theoretical Background

To compare the associative verbal network of the word and the test, Yu. Filippovich points out the need to construct a formal linguistic object and introduces the concept of stimulus-reactive chain, which is based on the statement: the same word-reactions of different respondents are equivalent. This, in turn, allows for each word-stimulus to build a frequency field of reactions and to present on this basis a chain of two or more stimulus-reactive pairs, ‘to continue’ or ‘to close’ stimulus-reactive chains, to present the relations between stimuli and reactions in the form of a grid of relations and graphs. For

example, WEAPON – SHOOTING; WEAPON – PISTOL – SHOOTING; WEAPON – RIFLES – SHOOTING; WEAPON – RIFLES – HUNTING – SHOOTING, WEAPON – ARMY – WEAPON – SHOOTING, WEAPON – ARMY – TORTURE – FIGHT – SHOOTING; WEAPON – BARREL – CANNON – CAST IRON – SHOOTING (a total of 141 chains); similar relations are built based on the expanded four-member syntactic constructions with different lengths of stimulus-reactive series like DUCKWEED IS A GREEN COVER OF RESERVOIRS, the graph shows different number of reactions to each of these components [8].

The associative test data are especially valuable for such an area as sentiment analysis, as they provide information on the axiology of the concept directly (scaling good /bad reactions and their taxonomic gradation) and indirectly through reactions characterized by certain evaluation. In addition, such AVN can be used in determining the gender of the author of the text.

It should be noted that, according to some linguists, “with any conceptualization of a fragment of reality, some aspects of reality are emphasized, actualized, others are obscured, go into the background: a schematization of reality takes place” [20].

The works of some linguists put forward the idea of creating a cognizer aimed at building a computational system to support empirical and theoretical cognitive research on verbal consciousness of a particular language personality, which includes computational modeling of mechanisms of language consciousness, focus on data from related disciplines, typology of knowledge, its volume and quality, experimental base, including cognitive experiments, which allow the theoretical development of the accumulated data structuring. The latter involves singling out the minimum unit (cognemes), larger units (concepts, superconcepts and conceptual domains) [21]. In the long run, this approach, in our opinion, is focused on typology of the specifics of sentiment analysis (it should be facilitated by the received responses aimed at emotional categorization of perception of various emotions and disloyal relationships in society, which include categories such as BETRAYAL, OFFENCE, ENVY, etc.) with their possible scaling and taxonomic grading involving modern interparadigmatic techniques, including conceptual semantic and ideographic analysis of negative emotions in phrasemes in paremiae of the languages of different structure using corpus data and other computational resources [22].

Methodologically, such fundamentally contrasting associative tests are carried out on the various thematic groups of vocabulary in closely related and structurally and geographically distant languages.

3 Methodology

This study presents a method of modeling the associative verbal network (hereinafter AVN) on the basis of the associative test data and, accordingly, reveals the principles of building an information system to present the axiological, socio-evaluative associative verbal network of speakers of different languages in typological and cognitive aspects.

This technique involves a comparative analysis of AVN of the studied languages: identifying the specificity of the associative content exemplified by Ukrainian and Slovak conceptual domains of ЗРАДА/ZRADA ‘BETRAYAL’, their taxonomy and typology of related frames.

Quantitative results are interpreted based on the taxonomy of frame structures and interconceptual associative relationships. It should be noted that the most important in this technique is to determine the ‘associative’ distance between concepts by analyzing data on their common associations (index of mutual associative relationship), as well as visualizing the results of the associative experiment, which allows us to identify such common areas. ‘Intuitive’ methods, in particular the introspection known in linguistics, should be verified by correctly interpreted quantitative data.

They are further visualized in the form of graphs and interframe associative complexes, with the help of computer programs and computer support based on axiological features, significantly supplemented by corpus data (IMAR, specificity of collocations), in particular, collocations, oppositional and interconceptual relations with other related domains and concepts, mental-synonymous /antonymous correlates.

The obtained results can be used not only in the formalization of new linguistic and computational techniques related to the study and formalization of psycholinguistic data, comparing the frequency ration of (absolute and relative) of reactions of respondents speaking different languages, but also in compiling new dictionaries of associative norms of the Slavs with a wide involvement of West Slavic content.

4 The Associative Verbal Network of the Ukrainian and Slovak Conceptual Domains ЗРАДА/ZRADA ‘BETRAYAL’

The associative test (free and intent) was carried out in 2019, involving 194 respondents – native speakers of the Ukrainian language; 134 respondents – speakers of the Slovak language, of different age groups and gender. In total, 67 stimuli were offered to the participants of the test.

The specificity of the proposed method is described on a fragment of AVN, as the data obtained in the Ukrainian language test alone include more than 12 thousand reactions.

This paper focuses on a fragment of AVN, which reflects the conceptual domain of Ukrainian ЗРАДА /Slovak ZRADA ‘BETRAYAL’.

Thus, disloyal actions, as figuratively betrayal is referred to by some Polish psychologists, always occur in society, where people betray each other in various spheres of social life – marriage, relationships between friends, subordinates and superiors, sellers and buyers, politicians and voters [23].

By profiling, according to the representatives of the Polish cognitive and textual school [24], we mean the taxonomic integrity of the facets arranged in a certain way forming a common gestalt and realized in the form of cognitive-semantic categorization of the described semantic object: {[SUBJECT] + [EVENT] + [PLACE] + [FUNCTION]} [25].

The graphic chart in Fig. 1 visualizes, based on the weight of each of the vertices, the associative verbal network of the Ukrainian concept of ЗРАДА ‘BETRAYAL’, which is verbalized by the units *зрада* ‘betrayal’, *зрадити* ‘to betray’, *зрадливий як* ‘unfaithful as (m)’, *зрадлива як* ‘unfaithful as (f)’, Fig. 2 shows AVN of the corresponding Slovak concept.

It is worth noting that the female and the male responses to the stimulus *зрада* ‘betrayal’ differ significantly. The common ones are presented in Fig. 3 based on the weight of each vertex.

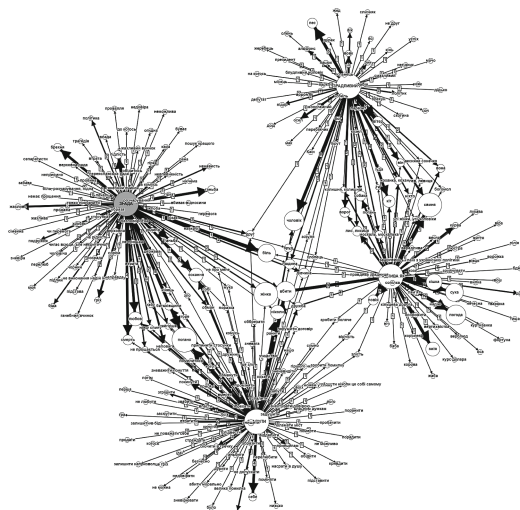


Fig. 1. The associative verbal network of the Ukrainian conceptual domain ЗРАДА ‘BETRAYAL’

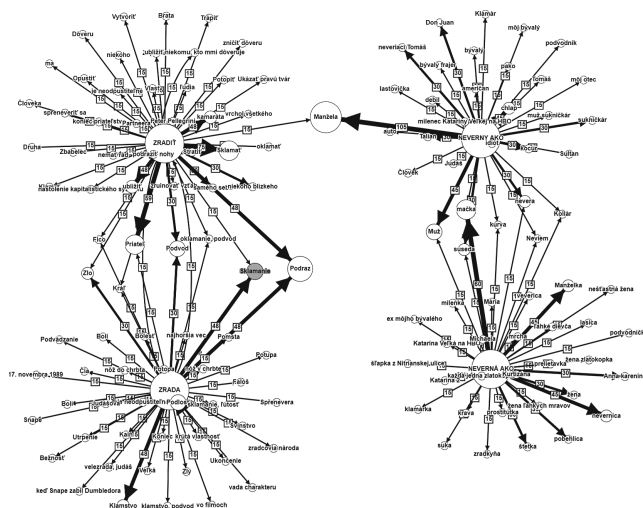


Fig. 2. The associative verbal network of the Slovak conceptual domain ZRADA ‘BETRAYAL’

The index of mutual associative relation (IMAR) of concepts is calculated by the ratio of the number of identical reactions to the total number of reactions received [26]. Thus, the IMAR for the Ukrainian concepts of ЗРАДА ‘BETRAYAL’ and ЗРАДИТИ ‘TO BETRAY’ is 0.4717. To compare, IMAR for ЗАЗДРИСТЬ ‘ENVY’ and ЗАЗДРИТИ ‘TO ENVY’ is 0.2140; for БІДА ‘MISERY’ and БІДУВАТИ ‘TO BE MISERABLE’ it is 0.040.

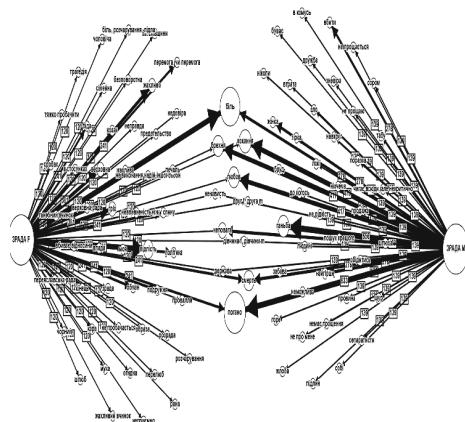


Fig. 3. Female and male responses to the stimulus *зрада* ‘betrayal’

The analysis of the results shows that the strongest associative relations are between the concepts of ЗРАДА ‘BETRAYAL’ and БРЕХНЯ ‘LIE’. This is confirmed by the conducted associative test and the obtained responses of both Ukrainians and Slovaks.

Yet, according to Word Similarity [27, 28], the most ‘similar’ to *зрада* ‘betrayal’ are the concepts of *помста* ‘revenge’, *обіцянка* ‘promise’, *ганьба* ‘disgrace’, *змова* ‘conspiracy’, *брехня* ‘lie’, *відвага* ‘courage’, *жадібність* ‘greed’, *провина* ‘fault’, *kráľovny* ‘queen’, *poéma* ‘poem’, *satirická* ‘satirical’, *krutá* ‘cruel’, *pomsta* ‘revenge’, *princezná* ‘princess’, *veselohra* ‘comedy’, *kliatba* ‘curse’.

A purely Slovak male reaction, to some extent unique, typologically distinct, not mentioned by Ukrainian respondents, is a hint at the mundaneness of betrayal, it is viewed as a commonplace: *bežnosť* ‘ordinary’ (*m* 5.88).

According to the GRAC [29] corpus (the frequency of the word *зрада* ‘betrayal’ is 29.30 uses per million), the most frequent collocations (except for a few grammatical ones) with the word under study are: *зрада* ‘betrayal’ – *подружній* ‘marital’, *державний* ‘state’, *батьківщина* ‘motherland’, *підлий* ‘mean’, *Батьківщина* ‘Motherland’, *національний* ‘national’, *Мазепин* ‘Mazepa’s’, *підступний* ‘insidious’, *Мазепа* ‘Mazepa’, *Юда* ‘Judah’, *зрадити* ‘to betray’ – *підло* ‘meanly’, *присяга* ‘oath’, *витримка* ‘endurance’, *Майдан/майдан* ‘Maidan/maidan’, *Батьківщина* ‘Motherland’, *таємниця* ‘mystery’, *підступно* ‘insidiously’, *могти* ‘can’, *ганебно* ‘disgracefully’, *віра* ‘faith’. Collocations are sorted by $MI.log_f$ ((formerly called salience) $MI-Score \cdot \ln(fAB + 1)$) [30]. The results obtained by means of this index, in our opinion, the best illustrate the concept of ЗРАДА ‘BETRAYAL’ and the most correlate with the results of the associative test, whereas T-score, MI and other formal statistical data fail to identify conceptual and axiological verbalizers of the concepts in the analyzed languages and semiotic spaces.

5 Taxonomy of Frame Structures: Comparative and Typological Aspect

Overall, the most frequent responses to the stimulus *зрада* 'betrayal' were as follows: *біль* 'pain' 5.16 (f7.23 /m2.78); *погано* 'badly' 5.16 (f 2.41 /8.33); *брехня* 'lie' 2.58 (f1.20 /m4.17); *ганьба* 'disgrace' 2.58 (f0.00 /5.56); *кохання* 'love' 2.58 (f1.20 /m 4.17); *любов* 'love' 2.58 (f 1.20 /m 4.17); *підлість* 'meanness' 2.58 (f4.82 /m 0.00); *смерть* 'death' 2.58 (f 2.41 /m 2.78). Concerning the female reactions, *біль* 'pain' (7.23); *підлість* 'meanness' (4.82); *верховна* 'supreme' (3.61); *гріх* 'sin' (3.61); *жахливо* 'terribly' (3.61) are of the highest frequency; regarding the male responses, such answers as *погано* 'badly' (8.33); *ганьба* 'disgrace' (5.56); *брехня* 'lie' (4.17); *кохання* 'love' (4.17); *любов* 'love' (4.17); *біль* 'pain' (2.78) are the most frequent.

In general, the Slovak nationally marked and specific reactions to the analyzed stimuli (ЗРАДИТИ 'TO BETRAY' and ЗРАДА 'BETRAYAL') include the following: *podvod* 'fraud', *potopa* 'flood', *podraziť nohy* 'to set foot', *kráľ* 'King', *krutá vlastnosť* 'cruel trait', *potupa* 'shame, disgrace', *vo filmoch* 'in films', *sprenevera*, *svinstvo* 'perfidy', *zbabelec* 'coward', *vrchol všetkého* 'the top of everything', *zhodiť* 'drop (from top to bottom)', *nastolenie kapitalistického systému* 'the establishment of the capitalist system', *Fico, Peter Pellegrini, November 17, 1989*. The explanation of the specifics of the latter responses, which belong to the reactions associated with *state (political) treason* and their value markedness are given in the corresponding subgroups of betrayal profiles.

Most of the answers were related to the moral evaluative conceptualization of reality, in other words, the respondents responded to the stimuli *зрада* 'betrayal', *зрадити* 'to betray' using expressive evaluations of the conceptual domain ЗРАДА 'BETRAYAL': *бруд* 'dirt' (0.61), *біда* 'misery' (1.29), *зло* 'evil' (1.26), *гіршого немає* 'nothing can be worse' (0.61), etc.

In addition, in the Slovak and Ukrainian languages associations with the destructive semantics are equally represented: Ukr. *поламати міст* 'to break the bridge', *порушити договір* 'to break the treaty', cf. Slovak. *zruinovať vzťah* 'to destroy relationships'.

Other metaphorical profiles of Slovak betrayal are confirmed by the following reactions: *potopa* 'flood', *potopiť* 'sink', mostly related to insidious, hostile actions to someone through the spatial-somatic code correlated with the idea of left and right (as correct, true): *podraziť* 'betray somebody', *podraziť nohy* 'trip someone' (2), *ukázať pravú tvár* 'to show one's true face' (2). The latter reaction can be attributed to the metaphorical model BETRAYAL IS A CAMOUFLAGE, DISGUISE, described in detail in the Polish language model of the world using the conceptual metaphor *treason - disguise - camouflage* [22].

Qualitative analysis of the obtained responses within the second profile on the basis of axiology showed the following: political state of affairs and authorities, social and political hierarchy of society, events in Eastern Ukraine, obviously, predominate in the answers of Ukrainian respondents, such as *батьківщини* 'of homeland', *батьківщина* 'homeland' (2.49), *верховна* 'supreme' (1.94), *верховна рада* 'supreme council' (0.65), *влада* 'authorities' (0.65), *державна* 'state' (1.29), *політика* 'policy' (1.29); *присязі* 'oath' (0.61), *сепаратисти* 'separatists' (0.65), etc. Concerning Slovak respondents, we observe isomorphic associative correlates: *velezrada* 'treason',

literally ‘great betrayal’, *sprenevera* ‘embezzlement’, *zradcovia národa* ‘traitor’, *vlast* ‘power’ (5.88).

Similar, but not very frequent reactions of Slovak respondents (mostly women) are focused on historical events, changes and transformations of the political system of society after 1989 or current statesmen, compare *nastolenie kapitalistického systému* ‘establishment of the capitalist system’ (m 2); *Fico* - this reaction is associated with a politician Robert Fico, who served as Prime Minister of Slovakia for 10 years, *Peter Pellegrini* is the name of the former Prime Minister; *November 17, 1989* is connected with the change of the political system of Slovak society, in particular the Velvet Revolution, which led to the overthrow of the communist regime.

Moreover, the stimuli *зрада* ‘betrayal’, *зрадити* ‘to betray’ received metaphorical responses in chronotope semiotic coordinates (space, locus) *навкруг* ‘around’ (0.65), *безповоротна* ‘irreversible’ (0.65), *кінець* ‘end’ (1.9) referring to the end of relationships. There are reactions related to spatial loci, which in turn objectify the emotional state, including the metaphor of the semiotic DOWN: *провалля* ‘abyss’ (0.65), while the stimulus ЗРАДИТИ ‘TO BETRAY’ correlates with the metaphorical expression of the idea of DOWN: *впасти* ‘to fall’ (0.61), *низько* ‘low’ (0.61).

Furthermore, there are a number of phraseological reactions: *ніж у спину* ‘a knife in the back’ (1.9); *впасти в очі* ‘fall in sb’s eyes’ (0.61), *спалити* ‘to burn’ (0.61), *скачати в/у гречку* ‘jump into buckwheat’ (1.22), *поламати міст* ‘break the bridge’ (0.61), the allusion to the saying *спалити/зруйнувати мости* ‘burn/destroy bridges’. An equivalent reaction can be found in the Slovak language mentality: *nôž do chrbta* ‘knife in the back’ (5.88). The benchmark response to the stimulus *зрадливий як* ‘unfaithful as (m)’ is *Janus*, which is associated with the biblical phrase *two-faced Janus*. The answer *Yarema Vyshnevetsky*, correlated with a historical figure, is nationally marked. The phraseological reactions of the Ukrainian males include *шкура* ‘skin’ (0.71) due to the phraseological convergence with the expression *продажна шкура* ‘corrupt skin’.

6 The Information System Modules for AVN Presentation

A. Filippovich proposed the principles of creating an automated system of scientific research on associative experiments, pointing out that this system should contain a number of subsystems: Electronic versions of associative dictionaries; System for conducting an interactive associative experiment; Associative verbal field analysis system; System for conducting a visual associative experiment; System of psycholinguistic analysis of texts [31]. In fact, the “automated system of scientific research on associative experiments” contains the following modules: Resources (Russian associative thesaurus (1.3 million records 1988–1995); Russian comparative associative dictionary; Associative dictionary of information technologies (12.5 thousand records). 2000); Slavic (Russian, Belarusian, Bulgarian, Ukrainian) associative dictionary; Dictionary of associative norms of the Russian language by A. A. Leontiev (25 thousand entries, 1967–1973); Module for searching for associative chains of the subsystem of the associative verbal field analysis; Subsystem for conducting an interactive associative experiment [32].

The experience of conducting associative experiments and further processing of the obtained data proves that the structural information system should contain the results

of previous associative tests, which will allow us to analyze the dynamics of AVN of Ukrainian native speakers. Thus, there are a number of associative dictionaries in Ukrainian psycholinguistics. "Dictionary of associative norms of the Ukrainian language" by N. Butenko was published in 1979 [33]. In 1989, N. Butenko's "Dictionary of Associative Definitions of Nouns in the Ukrainian Language" was published, which combines the idea of associative and attributive dictionaries. This dictionary is based on the results of AT with 200 respondents, who were given a list of 35-40 nouns, to each of which they had to write five to seven attributes (except for pronouns and ordinal numbers). The preface states that the most commonly used nouns of the Ukrainian language were used as stimuli [34]. In 2007, S. Martinek's "Ukrainian Associative Dictionary" was published [35]. The author used a list of 841 stimuli, "where words of different parts of speech are widely represented: nouns, adjectives, verbs, adverbs, etc. [35].

In addition, there are a number of 'specialized' associative dictionaries [36–38]. These data should make up one of the modules of the system (with the possibility to sort data from response to stimulus). "Associative Dictionary of Advertising Vocabulary" is one of a few works in the field of associative reflection of onymic units.

Therefore, the data accumulated in previous experiments should be the first module of the IS.

A questionnaire for conducting an associative experiment in real time should be the second module of the system. In addition, this module should be accompanied by a program for statistical analysis of data, such as: absolute and relative frequency of response to a particular stimulus; absolute and relative frequency of stimuli to which a certain response is given; possibility to sort data by age, gender, native language, education, profession, place of residence of the respondent, etc. Furthermore, as shown above, in order to interpret the results of the associative experiment, it is important to have such data as: the results of the classification of reactions by axiological feature, by type of reaction according to its type – paradigmatic/syntagmatic; determining the index of mutual associative relation.

Discussing the idea of the abovementioned cognizer, Yu. Filippovich emphasizes that for its creation in real time and introspection it is necessary to involve various concepts – living, relevant, obligatory (receptive), optional, personal, potential, professional, linguistic and cultural; at the same time, the software modeling subsystem of this module includes such components that perform semantic, operational, symbolic, subject, qualitative and evaluative, etc. functions that are essential for the associative portrait /image of the linguistic personality [8].

The third module of this system should perform the classification of the obtained reactions according to the corresponding profile and within the profiles of a frame or subframe, constructed on the basis of typology of stimulus-reactive and interconceptual relations (based on the relevance of vertices of represented graphs) by axiological, semiotic, gender and cultural characteristics.

Speaking about the structuring of AVN and analysis of linguistic consciousness in terms of stimulus-reactive mode of language consciousness, A. Filippovich emphasizes the need to involve the achievements of the related disciplines in this process, namely semiotics, psychology, culturology, sociology, aimed at creating an associative thesaurus, which is a prototype of the mentality of the 'average speaker' [31].

The fourth module is the representation of AVN in the form of a matrix and its visualization in the form of graphs based on the weight of the vertex and their semantic, associative-conceptual and figurative-metaphorical content in each compared linguistic worldview with clear explication of central and peripheral areas of AVN. The obtained data show that the latter include associative-symbolic and conceptual relations of the received reactions with related concepts – FAITHFULNESS-UNFAITHFULNESS, which require separate consideration (closer periphery of the associative verbal field), as well as interconceptual relations due to the dynamics of conceptual domain ЗПАДІА ‘BETRAYAL’ and “expansion of its ontology” (Yu. Filippovich’s term) (*deception, slander, envy, denunciation, conspiracy, revenge* and other one-order ideologemes and axiologemes, taken in terms of stimulus-reactive correlations), which constitute the further periphery of the associative field.

The fifth module of IS is a tool for comparing AVN of different languages in order to determine the common and distinctive features of the semantic-cognitive categorization of a particular conceptual domain of society, culture, ethnicity.

7 Conclusions

The proposed approach to identifying the features of cognitive experience of native speakers involved several stages of analysis with sequential quantitative parameterization of the data: total frequency (absolute and relative), frequency of reactions of a certain type by gender, age, etc. of native speakers; by axiological feature, by type of reaction considering its type – paradigmatic/syntagmatic; IMAR index. In addition, it is vital to compare the data of associative experiments with both preliminary results and corpus data. Therefore, the application aimed at automatic processing of statistical information is an important module of IS.

Therefore, the test data, which are summarized quantitatively and visually, are compared with the corresponding associative complexes, groups and networks of intersection of verbal reactions and their further modeling as frames. In addition, they are compared with mostly metaphorical collocations of ЗПАДІА ‘BETRAYAL’ taken from the GRAC corpus.

On the other hand, the research revealed that such psycho-cognitive, socio-cultural, semiotic and pragmatic phenomena as ЗПАДІА ‘BETRAYAL’ are defined by a complex taxonomy of conceptual-associative areas and/or conceptual domains, which is predetermined by the subject, object, participant, beneficiary, carrier, etc., which should be presented in the third module of the IS.

The visual representation (graphs 1, 2) of the areas of mutual associative relation of related conceptual domains as exemplified by the concept of ЗПАДІА ‘BETRAYAL’ provides qualitative and quantitative information about interconceptual relations, which allows the researcher to clearly illustrate the network nature of the mental level of human consciousness. Therefore, an advanced system must contain a module for AVN visualization.

As a result, the proposed approach made it possible to clearly present the facts of similarities and the existing differences between the Ukrainian and Slovak linguistic and mental, cultural, philosophical, ideological facts and models of worldview.

On the basis of the received reactions, their typology, and cognitive-statistical interpretation various semantic profiles of ЗПАДІА 'BETRAYAL' (marital, moral and ethical, national, military, loss and breach of trust, intrigue, scenario of CRIME AND PUNISHMENT) were received, which generally allows us to present figurative cognition of ЗПАДІА 'BETRAYAL' as a means of social interaction and cognitive interpretation of disloyalty to others (not oneself), reflected through the mental lexicon of Ukrainians and Slovaks, verified through the prism of applied, interparadigmatic and computational methods of information processing, storing and transmitting.

Thus, the creation of such an IS will allow us to introduce the Ukrainian language data into scientific circulation. To date, there is no digitized/electronic Ukrainian associative dictionary, although, as noted, there are a number of Ukrainian associative dictionaries compiled in different time periods, which in turn would make it possible to trace changes in the linguistic consciousness of native speakers. In addition, first of all, it is important for linguists to trace the so-called synchronous dynamics. In view of this, the second module of the system is proposed – a questionnaire for conducting an associative test in real time with statistical analysis of the obtained data. An important component of such a system is the automated/automatic classification of reactions on different grounds, which is possible on the basis of their vocabulary definitions. The effective analysis of the associative test results is achieved by means of visualization, which, in turn, facilitates interlingual comparison.

References

1. Horoshko, Y.: Interactive model of the free associative test. Kharkiv (2011)
2. Vasilieva, N.: Notes on psychoonomastics. Psycholinguist. Issues **14**, 128–137 (2014)
3. Zalevskaya, A.: Linguistic consciousness: questions of theory. Psycholinguist. Issues **1**, 30–34 (2003)
4. WordNet. <http://wordnet.princeton.edu>
5. GlobalWordNet. <http://globalwordnet.org>
6. Elenevskaya, M., Ovchinnikova, I.: Storage and description of verbal associations: dictionaries and thesauri. Psycholinguist. Issues **3** (29) (2016). <https://cyberleninka.ru/article/n/hrazenie-i-opisanie-verbalnyh-assotsiatsiy-slovari-i-tezaurusy>. Accessed 08 July 2020
7. Field, J.: Psycholinguistics: The Key Concepts. Routledge (2004)
8. Filippovich, Y.: Infocognitive technology for modeling verbal consciousness. NSU Bull. Linguist. Intercultural Commun. **10**(2), 51–62 (2012)
9. Basalkevych, O., Basalkevych, O.: Fuzzy simulation of historical associative thesaurus. Adv. Sci. Technol. Eng. Syst. **4**(5), 224–233 (2019)
10. Lyubymova, S.: Associative experiment in the study of a sociocultural stereotype. Kalby Studijos **36**, 85–96 (2020)
11. Panchenko, Alexander: Human and machine judgements for russian semantic relatedness. In: Ignatov, D., Khachay, M.Y., Labunets, V.G., Loukachevitch, N., Nikolenko, S.I., Panchenko, A., Savchenko, A.V., Vorontsov, K. (eds.) AIST 2016. CCIS, vol. 661, pp. 221–235. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52920-2_21
12. Kudryavtseva, E., Bubekova, L., Danilova, J.: Use of associative data dictionary for ethno-linguocultural interpretation of animated film. Media Watch **11**(1), 206–242 (2020)
13. Mukhametzyanova, L., Shayakhmetova, L.: Application of associative experiment in forming the foreign communicative competence. Engl. Lang. Teach. **7**(12), 60–64 (2014)

14. Morel Morel, D.: Comparing the same stimulus associative fields fixed in different historical periods: technique application case study. In: *New Paradigms and New Solutions in Modern Linguistics*, vol. 5, pp. 43–48, St. Petersburg (2014)
15. Bisikalo, O.: Construction of patterns chain within the model of associative pattern thinking. *Scientific Works of Vinnytsia National Technical University 2*, Vinnytsia (2009)
16. Ivanova, A., Yakushev, V., Timashova, L.: Visualization of the emotional vocabulary of the Chinese language as exemplified in the associative experiment. *XLinguae*. http://xlinguae.eu/files/XLinguae1_2018_21.pdf
17. Edinburgh Associative Thesaurus. <http://atour.iro.umontreal.ca/rali/?q=en/XML-EAT>
18. Russian associative dictionary. <http://www.thesaurus.ru/dict/>
19. Word Associations Network. <https://wordassociations.net/en/about>
20. Paducheva, E.: *Dynamic Models in the Semantics of Vocabulary*. Languages of Slavic culture, Moscow (2004)
21. Karaulov, Y.: Conceptography of the linguistic worldview. Article 1. The first stage of “ascent” to the image of the world: from elementary figures of knowledge to subject-reference areas of culture. *Scripta linguisticae applicatae*. Problems of applied linguistics 2, 7–17, Azbukovnik, Moscow (2004)
22. Tyshchenko, O.: Language means of ‘envy’ and ‘betrayal’ conceptualization: sphere of socially evaluating and emotional concepts and their interaction. In: Andreichuk, N., Babelyuk, O., Bialyk, V., Ivanchenko, M. (eds.) *Vectors of the development of philological sciences at the modern stage: collective monograph*. Liha-Pres, Lviv-Toruń (2019)
23. Błachnio, A.: *Psychologia zdrady*. Centrum Doradztwa informacji, Warszawa (2008)
24. Grzegorzczkowska, R.: Profilowanie a inne pojęcia opisujące hierarchiczną strukturę znaczenia. In: *Profilowanie w języku i tekście*, ss. 9–19, UMCS, Lublin (1998)
25. Bartmiński, J.: Dom – koncept uniwersalny i specyficzny kulturowo. In: Bartmiński, J., Bielińska-Gardziel, I., Żywicka, B. (eds.) *Leksykon aksjologiczny Słowian i ich sąsiadów*, vol. 1, Dom, pp. 15–33, UMCS, Lublin (2015)
26. Ulanovych, O.: *Psycholinguistics: a Coursebook*. Minsk (2010)
27. Word2Vec Models. <https://wordsimilarity.com/uk>
28. Word Similarity. <https://wordsimilarity.com/word-similarity-api>
29. Shvedova, M., Waldenfels, R., Yarygin, S., Kruk, M., Rysin, A., Starko, V., Woźniak, M.: General Regionally Annotated Corpus of Ukrainian (GRAC). Kyiv, Lviv, Yena (2017–2020). <http://www.uacorporus.org>
30. Statistics used in the Sketch Engine. <https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf>
31. Filippovich, A.: Automated system of scientific research on associative experiments. *Psycholinguist. Issues* 6 (2007). <https://cyberleninka.ru/article/n/avtomatizirovannaya-sistema-nauchnyh-issledovaniy-assotsiativnyh-eksperimentov-asni-ae>
32. Automated system of scientific research on associative experiments. http://it-claim.ru/Projects/ASIS/asni_grant.htm
33. Butenko, N.: *Dictionary of associative norms of the Ukrainian language*. Vyscha shkola (1979)
34. Suprun, A.: Foreword. In: Butenko, N.: *The dictionary of associative attributes of nouns in Ukrainian*. Vyscha shkola, Lviv (1989)
35. Martinek, S.: *Ukrainian Associative Dictionary: In 2 Volumes*. Ivan Franko National University of Lviv Publishing Center, Lviv (2007)
36. Kovalevska, T., Sologub, G., Stavchenko, O.: *Associative Dictionary of Ukrainian Advertising Vocabulary*. Astropoint, Odessa (2001)
37. Kutuza, N.: *Communicative suggestion in advertising discourse: psycholinguistic aspect*. Abstract of the dissertation for the degree of Doctor of Philology, I. Mechnikov National University of Odessa (2018)

38. Kovalevska, T.: Semantics of anonymous associations in advertising discourse. Notes on onomastics **5**, 3–11 (2001)

Artificial Intelilgence



Intelligent Neural Network Sensory System for the Analysis of Volatile Compounds in Beverages

Taras Chaikivskyi¹(✉)() ID, Bohdan Sus²(✉)() ID, Oleksandr Bauzha²(✉)() ID,
and Sergiy Zagorodnyuk²(✉)() ID

¹ Lviv Polytechnic National University, Bandera Str, 12, Lviv 79013, Ukraine
taras.v.chaikivskyi@lpnu.ua

² Taras Shevchenko National University of Kyiv, Volodymirskaya Str., 64, Kyiv 01033, Ukraine
bnsuse@gmail.com, asb@mail.univ.kiev.ua, kola@univ.net.ua

Abstract. An automated system for measuring the content of aromatic aldehydes in alcohol solutions has been developed. The main advantages of neural networks are compared with other mathematical methods, such as noise sustainability and the possibility of distributed data processing, the ability to process spectral dependencies in a wide range of measurements. An artificial neural network was created to process the output signals of the sensors, taking into account mutual cross-sensitivity and selective sensors to reduce the error of determining the concentration of volatile compounds. It has been shown that simple sensors can be integrated into an automated quality monitoring system for model vanillin mixtures. Simulation models were developed using sensors based on the electronic theory of sorption on the surface of semiconductors. The measuring complex can be adjusted to different measurement algorithms.

Keywords: Semiconductor sensor · Volatile solutions · Neural network · Microcontroller

1 Introduction

Characterization of volatile substances is an important task in the analytical chemistry, pharmaceutical and wine industries. Analysis of volatile compounds released by low-alcohol solutions can classify their specific characteristics, identify various problems that may occur during over-processing or long-term storage. Many components that affect the smell, taste and other properties of a beverage can alter the response of individual sensors and therefore discriminate against qualitative and quantitative criteria. Several methods make it possible to obtain basic characteristics of volatile substances. Each of these methods has its advantages and disadvantages. Common chemical analytical methods have high reliability, but they require long and complicated process of measuring samples [1].

The traditional analyzing of wine components is performed by chemical, physico-chemical and biochemical methods [2]. At the same time, the most modern approaches

were applied: high-performance liquid chromatography (HPLC) [3–5] and gas chromatography (GC) [5, 6], molecular weight chromatography [5], affinity chromatography using immobilized enzymes [7, 8], chemo- and biosensors with spectrophotometric [9, 10], chemiluminometric, fluorescence. methods of product registration, mass spectrometry (MS) [9], atomic and molecular adsorption spectrometry [10], including electrothermal, electron ionization (EI) [4]. To achieve the highest sensitivity and selectivity, several methods can be combined: HPLC /MS, GC /MS, HPLC /MS /MS, MS /MS /EI [11, 12].

Different variations of the concentration of volatile compounds and ethanol are mostly affecting to the aroma and taste of wine, respectively. Volatile compounds, mainly ethyl esters of organic acids, produce a “bouquet” of wine [13, 14]. The content of the ethyl esters (isoamyl acetate, ethyl hexanoate, ethyl octanoate, ethyl decanoate, etc.) in the wine increases during the aging process. In red wine, the sensory method of electronic semiconductor sensor array can distinguish and identify up to 800 different molecules of volatile compounds [15, 16]. Certain concentrations of protein, alcohol, and glycerol have been shown to significantly affect several notes of aroma. Most combinations have the significant effect with low concentrations of volatile compounds.

Wine contains antioxidants due to the presence of phenolic compounds. HPLC with spectrophotometric (SF) and fluorescence detection of the analyte revealed up to 48 different phenolic compounds in wine, namely anthocyanins, flavan-3-ols, flavanols, hydroxycinnamic and benzoic acids, etc. [17].

In the wine industry, product quality certification is used to control the quality and identification of wine flavors [18]. The quality of the wine is mainly evaluated by the means of physicochemical and sensory techniques. Nowadays, the use of semiconductor sensors is becoming increasingly popular. These systems, also known as “electronic tongues” and “electronic noses”, are based on different types of chemical sensors and biosensors with different transduction principles in combination with multi-factor data processing protocols [19, 20]. Parameters of volatile compounds in wine production are also monitored using multicomponent semiconductor gas analyzers. Semiconductor sensor systems provide simultaneous control of concentrations of several volatile substances. To increase information content, sensor systems are scalable, so the amount of data that needs processing as the number of sensors increases. Methods are also improved by using different measurement modes, data acquisition methods and mathematical processing algorithms. The sensitivity of the technique depends largely on mathematical software for computer processing of digital signals. Artificial neural networks are mathematical models developed on the principle of functioning of biological neural networks. The main features of these models are: learning ability, adaptability, noise resistance and fault tolerance [21].

Due to these features, the neural network method has been successfully applied in the fields of optimization of solutions, forecasting, classification, pattern recognition, control and filtering of data. In natural sciences, neural networks are used for approximation of functions, modeling of physical processes, and recognition of peaks in spectral dependencies [22].

The progress demonstrated by neural networks, as software operating under traditional operating systems and computer architectures, substantiated the relevance of

the development of neurocomputers and specialized microcontrollers. The system-based implementation of these devices corresponds to the ideology of neural network functioning.

To increase the performance of the neural network during solving a complex classification problem, it is advisable to separate the initial complex task into a set of classification tasks with reduced complexity. The number of neurons in the input, output and hidden layers can be used as indicators of complexity.

The complexity can be reduced at both the problem-solving level and the processing procedure level. For example, in Tree Divide To Simplify -T-DTS, the task is subdivided into subtasks recursively and the computational structure of the neural tree is generated. The splitting is performed by a mechanism of complexity estimation, which acts as a cycle of regulation on the decomposition process [23].

The learning process consists of decomposing data and storing processing of sub-structures and tools for decomposed datasets. Along with the principles of functioning of the model of a neuron as an elementary node of an artificial neural network and features of architecture, a decisive factor that ensures the efficiency of the use of neural networks is the teaching method. Teacher training method is based on the use of a sequence of training samples, and the learning mechanism consists in the modification of neuronal weights. The most common method for such neural network training is the gradient descent algorithm [24].

However, the relatively high measurement error of the concentrations restrains their scope. High-selective sensor systems are promising to reduce the error associated with the cross-sensitivity of gas analyzers. They are an important part of the Lab_on_a Chip lab concept and allow for qualitative or quantitative analysis in near real-time [25]. Semiconductor-type gas-sensitive sensors have low response times, but they are also sensitive to parameters such as light, temperature, humidity, etc. The effect of these factors has a significant effect on the measuring signal. Currently, a number of technological approaches are being used to minimize the impact of these factors. Preferably, these are the methods of temperature compensation of gas-sensitive sensors and the impulse heating mode of the sensors. The effect of evaporation of water and ethanol components on the volatile mixture is greatly reduced in the works [26]. For this purpose, the measuring cell was purged for 10 additional minutes with helium. It should be noted that technologically such a procedure is quite expensive. Such solutions do not always provide the required level of error due to the effects of temperature and humidity. Besides, it increases the reaction time of the gas analyzers.

It has been shown that under conditions of light illumination, the photocurrent of a semiconductor greatly depends on the recombination characteristics and the charge state of the surface. The approach of using solar cells as sensory structures is interesting. Induced light beam (LBIC) measurements reflect the spatial distribution of the solar cell photocurrent. Depending on the diameter of the excitation light beam and the distance, a lateral resolution of several hundred microns to several microns can be achieved. The concept of the LBIC sensor system is described in [27]. The conductivity of the sensitive layer of a semiconductor sensor depends on the concentration of free electrons, which is comparable to the fraction of the surface occupied by the adsorbed volatile compound.

However, such sensors also have low selectivity and significant measurement error. The responses of each low-selective sensor are slightly dependent on the type of adsorbent. An option for increasing selectivity is to increase the number of sensors and create their array. The set of multiple responses to the sensory system creates a unique imprint of substance. To detect the composition of volatile compounds in sensor systems, an artificial neural network apparatus can be used. The computer evaluates the signal pattern and compares the flavors of different samples. These results are comparative and can be presented as a “imprint”.

In particular, in [28] the advantage of machine learning techniques such as linear regression, neural network and vector machine support to determine the dependence of wine quality on other variables and predict wine quality is investigated. It has been shown that the value of a dependent variable can be more precisely predicted if only the important features are taken into account in the forecasting and not all the features are taken into account. Linear regression was implemented to determine the dependence of wine quality on various 11 physicochemical characteristics. Wine quality assessment is one of the key elements in this context and this evaluation can be used for certification. This type of quality certification helps to ensure the quality of the wine on the market. The wine has different characteristics such as density, pH, alcohol and other acids. Low-accuracy electronic sensor systems use pattern recognition to gain selectivity. Artificial neural networks can also be effectively implemented to increase the selectivity of semiconductor sensor systems.

Producing multi-component gas analyzers based on neural networks that give simultaneous quantitative analysis to determine component concentrations, increase selectivity and reduce sensitivity to environmental factors is now a relevant task.

Since, specialists in physics, chemistry and electronics were involved in research, the project included:

- studies of absorption, surface recombination and change of photocurrent in Silicon barrier structures and ways to optimize photoelectric conversion in sensors manufactured on these structures.
- developed structural and hardware circuits of the measuring installation.
- initial studies of the content of volatile components in alcoholic beverages depending on the shelf life, created model mixtures.
- developed and configured a neural network to process data obtained from the experiment. The obtained experimental results were statistically verified.

2 Hardware Implementation

A block diagram of a multicomponent volatile analyzer and a measuring chamber are shown in Fig. 1 and 2, respectively.

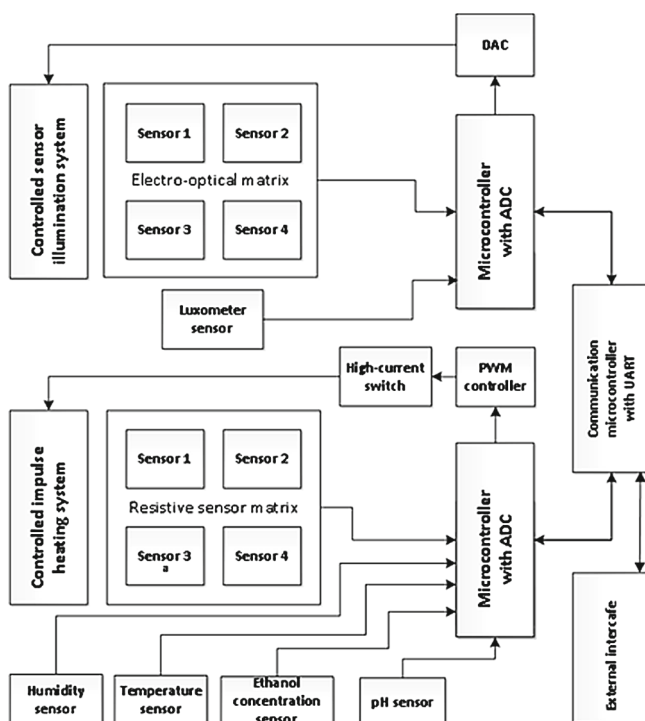


Fig. 1. Block diagram of a multicomponent measurement complex.

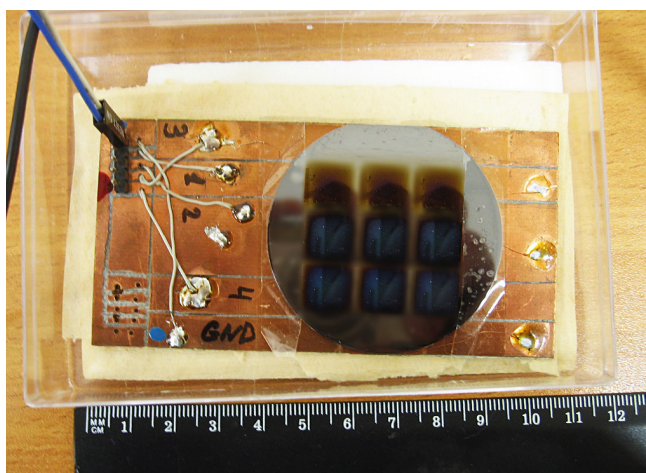


Fig. 2. Optical conversion sensor measuring camera.

Now, the most promising are touch systems based on digital electronic devices - microcontrollers. The system consists of sensors that are sensitive to volatile compounds

and sensors of environmental parameters, microcontrollers of control and preprocessing of signals and microcontroller of communication. A separate microcontroller controls the illumination of the touch panel illumination in the specified range. Brightness values are received from the illumination sensor. The brightness coefficients and the corresponding photoresponses of different samples are memorized by the neural network. The communication microcontroller receives data from ADC microcontrollers and processes the data according to algorithms implemented by the artificial neural network. ADC microcontrollers convert analog sensor signals into digital signals, perform signal correction and normalization, control sensor backlight modes with optical conversion and change in temperature measurement modes. Additional software and hardware encryption may also be performed [29, 30]. The developed device can optionally work with wireless data modules and be used for laboratory work [31, 32].

While using a multi-barrier sensor with a barrier structure, the p-n junction is fed by a bias voltage. Each channel has a distinct sensitivity to analytes. The equivalent circuit of the sensor is shown in Fig. 3.

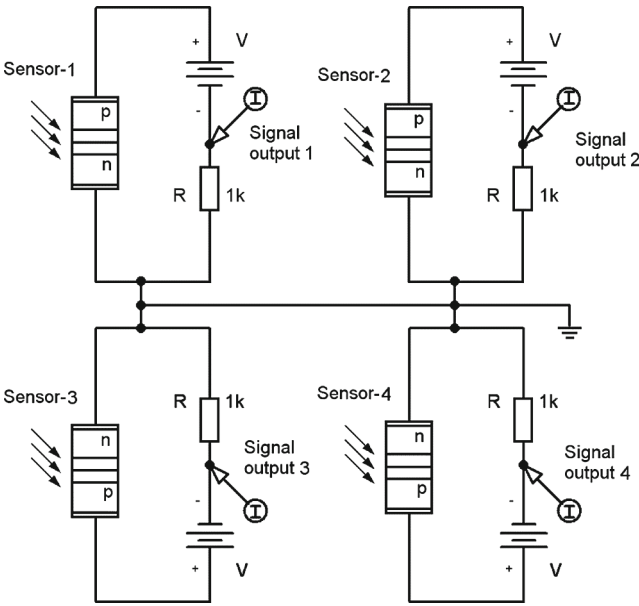


Fig. 3. Equivalent circuit of 4-channel sensor with p-n junction.

Sensor values also depend on the illumination brightness of the p-n junctions. The DAC controls the change in brightness in an addressed range during measurements.

3 The Experiment

To evaluate the experimental data for calibration of sensors, reference solutions of vanillin in 40% ethyl alcohol were collected. The results of measurements of the intensity of volatile compounds obtained from real samples of alcoholic beverages can be compared with the reference samples.

Vanillin is one of the main cognac markers responsible for its authenticity. The amount of vanillin in cognac is directly proportional to its age, the average rate of accumulation of vanillin is 0.57 mg/l per year [33]. For studies were used solutions with vanillin concentration 1, 1.25, 1.5, and 2 mg/l in 40% ethanol, which corresponds to cognacs for 3–7 years old.

The test specimens were placed under electronic nose sensors. The sensors were illuminated with light of equal intensity. The results of photocurrent measurements from four electronic nose sensors are shown in Fig. 4.

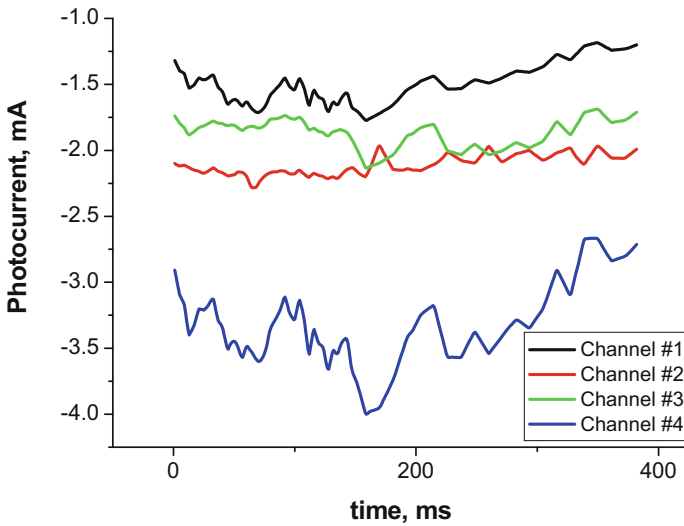


Fig. 4. Results of photocurrent measurements from the sample #4.

The dependences of the sensor readings from different samples are shown in Fig. 5. The X-axis represents the normalized values of the photocurrent. Along the Y-axis, the density of currents in the sample of measurements is shown. Photocurrents from different samples are shown in different colors. As can be observed from the Fig. 4., the values on the sensors from different samples overlap.

Due to the fact that the results of the sensor performance for the used cognac samples have similar values and no clear correlation of the sensor readings from the concentration of vanillin in the samples was observed, it was decided to use an artificial neural network for substance classification. The architecture of the neural network is a multilayer neural network of direct propagation with full filling of the connections.

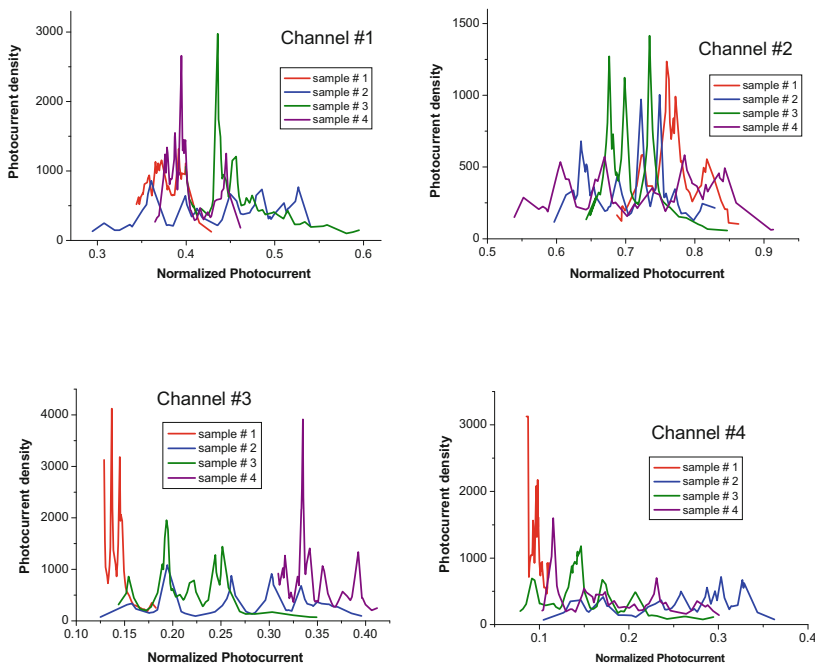


Fig. 5. Errors of measurements.

4 Software Implementation

Teacher training is based on the use of a series of training samples that set the desired values of the output vectors of the neural network for the corresponding values of the input vectors. The mechanism of learning consists in the modifying the weights of neurons. The main model for evaluating performance-based learning is the objective function, which provides a functional description of how the model output matches the ideal one for a given input sample. The target function for teaching methods with the teacher is also called the error of the original vector of the artificial neural network.

The mean square objective function is used:

$$E = \frac{1}{2N} \sum_{i=1}^N [\varphi_w(x^{(i)}) - y^{(i)}]^2$$

where N is the number of pairs of input and output vectors (frames) in the training sample; $x^{(i)}$, $y^{(i)}$ pair of input and output vectors; φ_w approximating function.

The sigmoid was taken as the activation function

$$f(x) = \frac{1}{1 + e^{-x}}$$

The artificial neural network receives sets of values of a variable, equivalent concentrations of the mixture of volatile compounds from optical and resistive sensor matrices.

The input layer of the neural network consists of 8 input neurons. In this experiment, data from photocurrent sensors were taken from 4 channels. Temperature and humidity during the experiment were kept in close range and were not used as input parameters of the system. The values of the input signals have been normalized to 1. The normalized values of the photocurrents were the input values for the neural network. The neural network had three hidden layers with 10 neurons each. The source layer had 6 neurons (source neurons are responsible for a specific substance). The structure of the neuron network is shown in Fig. 6. This artificial network allows refining the required number of neurons in layers.

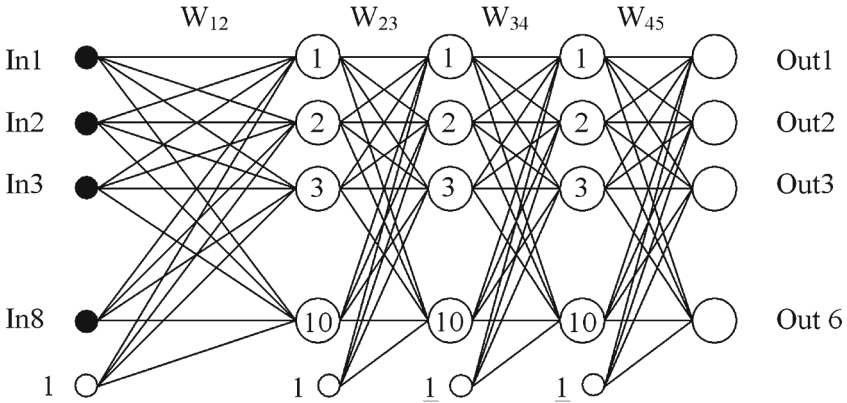


Fig. 6. Multilayer feedforward artificial neural network.

Artificial neural network training was performed by a method of teacher training. To find the minimum of the error function, the method of the inverse error propagation in stochastic gradient descent algorithm was implemented. In case the neural network could not immediately determine the composition of the volatile components from the data set, it begins to search for the closest connection. The algorithm of training artificial neural network is shown in Fig. 7. Where **In** and **Out** are the input and output vectors of the artificial neural network, respectively. **W** - matrix of weight coefficients of the learning coefficients of the artificial neural network.

The user interface of the sensor software with the neural network implementation is shown in Fig. 8.

The “New System” button allows the user to configure the neural network. During configuration, it is possible to specify the number of input, output neurons and the number of neurons in each of the 3 intermediate layers. The “Set I/O”, “Save I/O”, “Set Value”, and “File I/O” buttons define the input vectors of the input and output neurons (frames). The process of training the neural network begins with the “Train” button. The Empty Weights button sets the weight of the W_{ij} training to its initial, random position. The “Initialize Weights” button changes the learning scales and exits the system from the point of local minimum error.

The values of the input and output neurons in the frame are represented at the top left in the form of a matrix of rectangles. The color of the rectangles is determined by

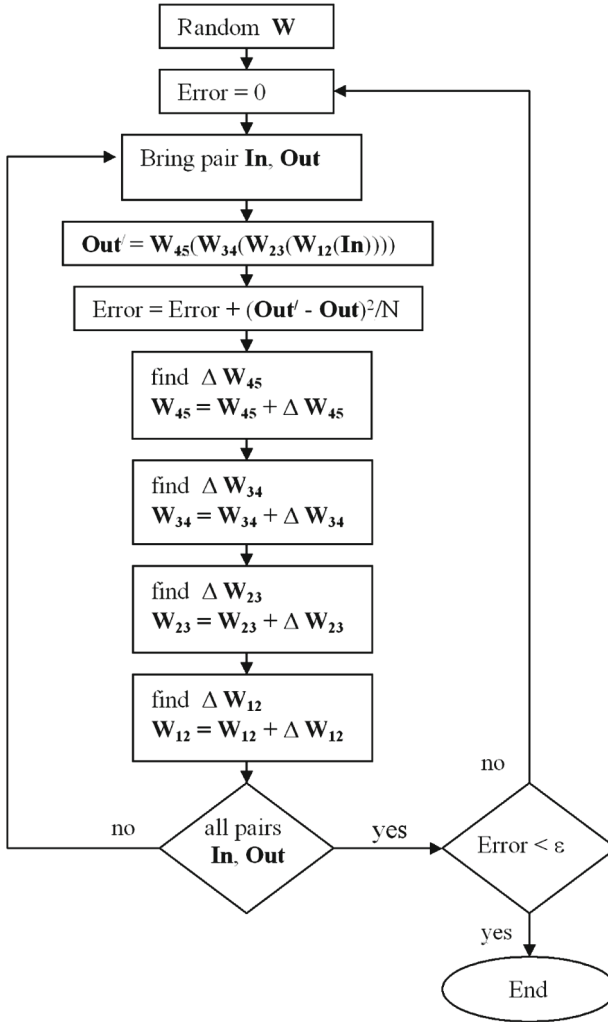


Fig. 7. Backpropagation algorithm.

the value of the corresponding neurons. The scale for matching and changing the color and meaning of the neurons is shown at the bottom right.

In Fig. 8 it is shown that the neural network has completed the training and the schedule of the error change can be observed below.

If a local minimum occurred in the neural network as a result of examining for the global minimum of error, the operator had the opportunity to randomly change (up to 10%) the weights. That led to the system dropping the local minimum, and the neural network continued to search for the global minimum of error during further training.

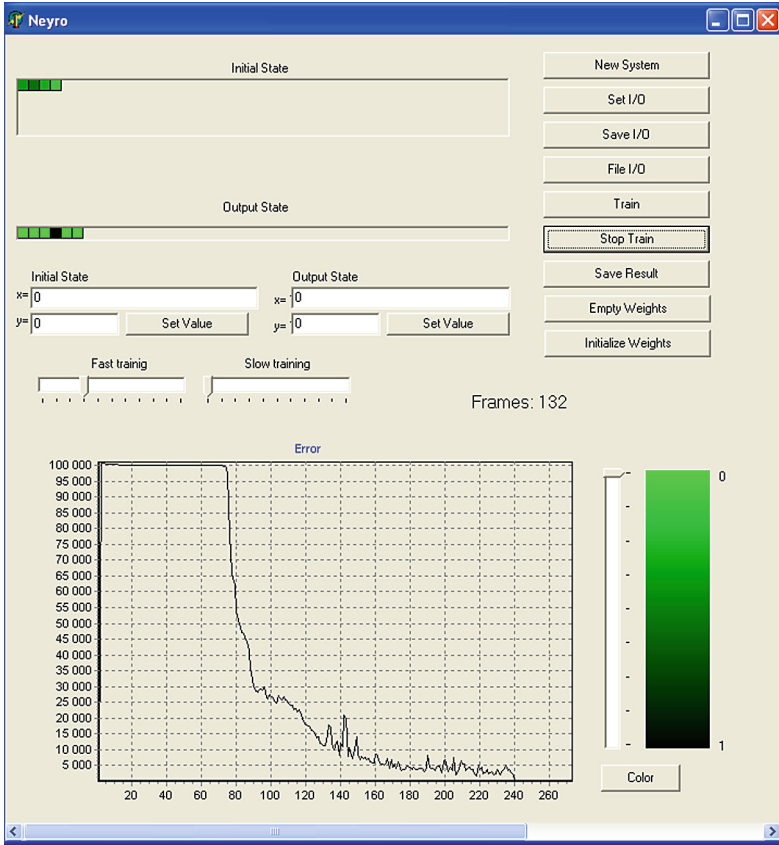


Fig. 8. Software interface.

5 Measured Signal Processing

From the experimental results, two samples of input values for the neural network were taken. Normalized photocurrent values were submitted to the inputs of the neural network; the Winner in the artificial neural network had to be one output neuron responsible for the given odor.

The first sample consisted of 132 sets of sensor displays and formed a selection of frames for artificial network training. The second sample consisted of 72 sets of sensor indicators and served to test the system for the ability of the trained neural network to distinguish non-training sample data. The training involved 4 volatile compounds. Each compound had the same number of input sets in both the training and control samples.

During training, the neural network reached an error of $\varepsilon < 0.01$ in approximately 400 learning periods. At the end of the training, the weights of the artificial neural network were recorded and the results of the training were checked with a control sample. In 98.3% of cases, the neural network reported the correct result. In 1.7%, the winner was the neuron that did not correspond the original odor data set. However, in

35% of cases, the winner was a neuron that did not match the authentic odor of the others with indicators of 3–44% was the neuron that matched the authentic odor. In the rest of the cases (65%), the winner was unconditional, and the rest of the original neurons had scores less than 2%.

Further learning of the error $\varepsilon < 0.001$ resulted in the following changes:

- In 0.1% of cases when the neural network indicated the correct odor, the definition of the leader changed to the wrong one after the training (error reduction).
- In all 1.7% of cases where the original neuron responsible for the non-authentic odor was victorious, the winner remained existing after retraining (error reduction). However, the share of second place neurons decreased. So a worthy second place with an error $\varepsilon < 0.001$ in 12% of cases, in the victory of a neuron with a non-authentic odor was occupied by a neuron corresponding to the authentic odor with indicators 3–42%. The proportion of the original neuron corresponding to the authentic odor can both decrease and increase.

6 Conclusions

The suggested setup enables to carry out the primary analysis of cognacs (or aged strong alcoholic beverages) for the presence of vanillin peak. In a case of peak missing, or when the content of vanillin does not fit into the typical range for beverages of the appropriate age of aging, this sample is dropping out. If the preliminary analysis is positive, then a detailed analysis of the quality of the drink should be performed by chemical chromatography. In addition to the vanillin peak, there are several other phenolic aldehydes that are considered as quality markers.

As the number of test substance types increases, the number of selective sensors to monitor substance types should be also increased. Despite this, system should be used for rapid preceding analysis.

The proposed approach provides additional data volumes and enhances component identification accuracy.

Processes that take place on the surface of a sensitive semiconductor layer when the temperature, light, and humidity of the environment are taken into account

Measurement errors in the determination of concentrations of volatile mixtures in multicomponent sensors when exposed to external disturbances are reduced.

Mathematical models of the transformation processes in the sensors were used, which were used in the training of neural networks. By the means of a neural network, sensors can be calibrated to increase selectivity and to predict complex integral features in the samples.

When considering a substance different from the neural network, one should take into account not only the winners from the class of source neurons but also the source neurons that received second place.

Despite the neural network architecture is simple and powerful, the increase in the number of samples leads to an increase in both the complexity of the neural network and the number of sensors needed to accurately classify the test substances.

Some substances may have a range of odors that can be confusing to the neural network, so the proper classification of substances by component is an important task that needs more detail consideration.

The main areas of our further research are the search for methods for establishing the age of beverage, namely, the justification for the use of markers and their correlation to improve the reliability of results.

References

1. Waterhouse, A.L., Ebeler, S.E. (eds.) Chemistry of wine flavor. American Chemical Society; Distributed by Oxford University Press, Washington, DC (1998)
2. Jackson, R.S. (ed.) Speciality wines. Elsevier, Acad. Press, Amsterdam (2011)
3. Jackson, R.S.: Wine science principles and applications. Elsevier Acad. Press, Amsterdam (2008)
4. Restani, P., Uberti, F., Tarantino, C., Ballabio, C., Gombac, F., Bastiani, E., Bolognini, L., Pavanello, F., Danzi, R.: Validation by a collaborative interlaboratory study of an ELISA method for the detection of caseinate used as a fining agent in wine. *Food Anal. Meth.* **5**, 480–486 (2012). <https://doi.org/10.1007/s12161-011-9270-9>
5. Marsili, R. (ed.) Flavor, Fragrance, and Odor Analysis, CRC Press, Cambridge (2001). <https://doi.org/10.1201/9780203908273>
6. Jackson, R.S.: Wine tasting: A Professional Handbook. Elsevier/Academic Press is an imprint of Elsevier, Amsterdam (2017)
7. Uthurry, C.A., Lepe, J.A.S., Lombardero, J., García Del Hierro, J.R.: Ethyl carbamate production by selected yeasts and lactic acid bacteria in red wine. *Food Chem.* **94**, 262–270 (2006). <https://doi.org/10.1016/j.foodchem.2004.11.017>
8. Mira de Orduña, R., Liu, S.-Q., Patchett, M.L., Pilone, G.J.: Ethyl carbamate precursor citrulline formation from arginine degradation by malolactic wine lactic acid bacteria. *FEMS Microbiol. Lett.* **183**, 31–35 (2000). <https://doi.org/10.1111/j.1574-6968.2000.tb08929.x>
9. Karadjova, I., Izgi, B., Gucer, S.: Fractionation and speciation of Cu, Zn and Fe in wine samples by atomic absorption spectrometry. *Spectrochim. Acta Part B* **57**, 581–590 (2002). [https://doi.org/10.1016/S0584-8547\(01\)00386-X](https://doi.org/10.1016/S0584-8547(01)00386-X)
10. Ajtony, Z., Szoboszlai, N., Suskó, E.K., Mezei, P., György, K., Bencs, L.: Direct sample introduction of wines in graphite furnace atomic absorption spectrometry for the simultaneous determination of arsenic, cadmium, copper and lead content. *Talanta* **76**, 627–634 (2008). <https://doi.org/10.1016/j.talanta.2008.04.014>
11. Pan, X.-D., Tang, J., Chen, Q., Wu, P.-G., Han, J.-L.: Evaluation of direct sampling method for trace elements analysis in Chinese rice wine by ICP–OES. *Euro. Food Res. Technol.* **236**, 531–535 (2013). <https://doi.org/10.1007/s00217-012-1888-3>
12. Jiao, Z., Dong, Y., Chen, Q.: Ethyl carbamate in fermented beverages: presence, analytical chemistry, formation mechanism, and mitigation proposals: ethyl carbamate in fermented beverages.... *Compr. Rev. Food Sci. Food Saf.* **13**, 611–626 (2014). <https://doi.org/10.1111/1541-4337.12084>
13. Villamor, R.R., Evans, M.A., Mattinson, D.S., Ross, C.F.: Effects of ethanol, tannin and fructose on the headspace concentration and potential sensory significance of odorants in a model wine. *Food Res. Int.* **50**, 38–45 (2013). <https://doi.org/10.1016/j.foodres.2012.09.037>
14. Muñoz-González, C., Rodríguez-Bencomo, J.J., Moreno-Arribas, M.V., Pozo-Bayón, M.Á.: Beyond the characterization of wine aroma compounds: looking for analytical approaches in trying to understand aroma perception during wine consumption. *Anal. Bioanal. Chem.* **401**, 1497–1512 (2011). <https://doi.org/10.1007/s00216-011-5078-0>

15. Franc, C., David, F., de Revel, G.: Multi-residue off-flavour profiling in wine using stir bar sorptive extraction–thermal desorption–gas chromatography–mass spectrometry. *J. Chromatogr. A* **1216**, 3318–3327 (2009). <https://doi.org/10.1016/j.chroma.2009.01.103>
16. Ragazzosanchez, J., Chaliel, P., Chevalier, D., Calderonsantoyo, M., Ghommidh, C.: Identification of different alcoholic beverages by electronic nose coupled to GC. *Sens. Actuators B: Chemical* **134**, 43–48 (2008). <https://doi.org/10.1016/j.snb.2008.04.006>
17. Gómez-Alonso, S., García-Romero, E., Hermosín-Gutiérrez, I.: HPLC analysis of diverse grape and wine phenolics using direct injection and multidetection by DAD and fluorescence. *J. Food Compos. Anal.* **20**, 618–626 (2007). <https://doi.org/10.1016/j.jfca.2007.03.002>
18. Macías, M., Manso, A., Orellana, C., Velasco, H., Caballero, R., Chamizo, J.: Acetic acid detection threshold in synthetic wine samples of a portable electronic nose. *Sensors* **13**, 208–220 (2012). <https://doi.org/10.3390/s130100208>
19. Lvova, L., Kirsanov, D.: Multisensor systems for analysis of liquids and gases: trends and developments. *Front. Chem.* **6**, 591 (2018). <https://doi.org/10.3389/fchem.2018.00591>
20. Oliinyk, B.V., Isaieva, K., Manilov, A.I., Nychporuk, T., Geloen, A., Joffre, F., Skryshevsky, V.A., Litvinenko, S.V., Lysenko, V.: Silicon-based optoelectronic tongue for label-free and nonspecific recognition of vegetable oils. *ACS Omega* **5**, 5638–5642 (2020). <https://doi.org/10.1021/acsomega.9b03196>
21. Rumelhart, D.E., McClelland, J.L.: University of California, S.D., PDP Research Group: Parallel distributed processing, explorations in the microstructure of cognition (1987)
22. Chow, T.W.S., Cho, S.-Y.: Neural networks and computing: learning algorithms and applications. Imperial College Press (2007). <https://doi.org/10.1142/p487>
23. Bouyoucef, E., Chebira, A., Rybnik, M., Madani, K.: Multiple neural network model generator with complexity estimation and self-organization abilities. *Int. Sci. J. Comput.* **4**(3), 20–29. ISSN 1727-6209 (2005)
24. Haykin, S.S.: Neural Networks: A Comprehensive Foundation. Prentice Hall, Upper Saddle River (1999)
25. Ghafar-Zadeh, E., Sawan, M.: CMOS Capacitive Sensors for Lab-on-Chip Applications: A Multidisciplinary Approach. Springer, Dordrecht (2010)
26. Guadarrama, A., Fernández, J.A., Íñiguez, M., Souto, J., de Saja, J.A.: Array of conducting polymer sensors for the characterisation of wines. *Anal. Chim. Acta* **411**, 193–200 (2000). [https://doi.org/10.1016/S0003-2670\(00\)00769-8](https://doi.org/10.1016/S0003-2670(00)00769-8)
27. Litvinenko, S.V., Kozinetz, A.V., Skryshevsky, V.A.: Concept of photovoltaic transducer on a base of modified p–n junction solar cell. *Sens. Actuators A* **224**, 30–35 (2015). <https://doi.org/10.1016/j.sna.2015.01.014>
28. Gupta, Y.: Selection of important features and predicting wine quality using machine learning techniques. *Procedia Comput. Sci.* **125**, 305–312 (2018). <https://doi.org/10.1016/j.procs.2017.12.041>
29. Tmienova, N., Sus, B.: Hardware data encryption complex based on programmable micro-controllers. In: CEUR Workshop Proceedings, pp. 199–208 (2018). <http://www.ceur-ws.org/Vol-2318/paper17.pdf>
30. Bauzha, O., Sus, B., Zagorodnyuk, S., Stuchynska, N.: Electrocardiogram measurement complex based on microcontrollers and wireless networks. In: International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, PIC S and T, pp. 345–349 (2019)
31. Sus, B., Tmienova, N., Revenchuk, I., Vialkova, V.: Development of virtual laboratory works for technical and computer sciences. In: Damaševičius, R., Vasiljevičienė, G. (eds.) Information and Software Technologies, pp. 383–394. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30275-7_29

32. Chaikivskyi, T., Bauzha, O., Sus, B. B., Tmienova, N., Zagorodnyuk, S.: 3D simulation of virtual laboratory on electron microscopy. In: CEUR Workshop Proceedings 2533, pp. 282–291 (2019). <http://ceur-ws.org/Vol-2533/paper26.pdf>
33. Chaikivskyi, T., Sus, B., Hunkalo, A.: Microcontroller-based multi-channel sensor system for monitoring the quality of spirit beverages. In: 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), pp. 59–63 (2020) <https://doi.org/10.1109/TCSET49122.2020.235391>



AI System in Monitoring of Emotional State of a Student with Autism

Vasyl Andrunyk^(✉)  and Olesia Yaloveha

Lviv Polytechnic National University, Lviv, Ukraine
{vasyl.a.andrunyk, olesia.yaloveha.mnpz.2020}@lpnu.ua

Abstract. This article is an overview of a practical implementation of monitoring the emotional state of a student with autism. In this article, we present an in-depth review of current technologies for detecting human emotions. The purpose of this work is to create an AI system to help teachers and psychologists to observe children with autism while they are learning or passing tests. The object is the methods and tools for emotion recognition. This system is designed to solve the problems of recognizing the emotional state of the student and helping him to properly respond to the emotional manifestations of other people. The subject of this system are processes of activity of this AI. We conclude the investigation by highlighting the aspects that require further research and development.

Keywords: Information system · Emotion · Recognition · Artificial intelligence

1 Introduction

Emotions are an important component of human communication and interaction. We often rely on them both in everyday life and in unusual situations. They can be expressed in various ways: facial expressions, posture, movements, voice, body response (heart rate, blood pressure, respiratory rate). However, the best reference is the human face [1–3].

There is a false statement that people with different nosologies are people who have an emotional deficit. As an example, many people with autism simply do not express emotions in the way that average people show them. That is, the idea that people with autism tend to lack empathy and cannot recognize their feelings is wrong [4].

The founder of the science of emotions and their recognition is the American psychologist Paul Ekman. He defined the classification of basic emotions of man, which are still guided in the 70s of the last century. This includes six emotions: anger, disgust, fear, happiness, sadness and surprise [5, 6].

A global breakthrough in the development of emotion recognition technology has come not too long ago. He has created various applications and algorithms that can determine the emotional state of the user. For example, the Text Analytics API is one of the Microsoft Cognitive Services, which allows developers to embed ready-made “smart” algorithms into their products. Already existing software tools often work with neural networks in real time. This allows us to apply software in various areas of our lives and thus actively influence its quality [7, 8].

Emotional recognition using facial expressions (ERFE) is of great interest, despite advances in artificial intelligence, mainly due to the large variability in the use of technology in various areas of human life, such as medicine, video games, security, human-computer interaction and affective computing. The human face includes most of the information about human sensations, and facial expressions are the main path used to transmit emotions. These facts underline the importance of facial expressions for recognizing emotions and substantiate the interest that ERFE has generated over the past two decades.

As software that supports 3D technology, 3D ERFE algorithms are largely model-based and popular for identifying basic geometric features. Ekman and Friesen [8] (1978) proposed the Facial Encoding System (FACS), which was the first comprehensive technology for recognizing all visually distinct, noticeable facial movements called action units. Many 3Dface databases have been created based on FACS and some other models containing expression data such as Bosphorus, ICT-3DRFE and D3DFACS. They use professional 3D equipment with high resolution but low scanning speed. Advanced 3D imaging equipment has become increasingly popular in the ERFE field, including the Kinect sensor (Microsoft, USA) and the Creative Interactive Gesture Camera (CIGC).

Today, less attention is given to identifying the emotions that children with ASD exhibit than their ability to identify other people's facial expressions. This aspect is important for the productive and fruitful learning of children with autism, because if interpreted emotionally correctly, it will help to modernize the learning process according to the needs of the students. The use of emotion recognition technology can make this process quite significant [9–14].

2 Description of the System

The purpose of this study is to identify the emotional state of the student as they pass lessons/tests to correct further treatment. Because inclusive classes only have a teacher and a teaching assistant, it is impossible for them to concentrate their full attention on one student, even if there are only two students. As children with autism are very sensitive to the environment, they immediately notice changes in attitudes of adults [15]. It is also impossible to just send a few people out to track the emotional state of the students because the children simply will not allow strangers to be near them and it will take a long time to earn the trust of the children. Therefore, a system for detecting a student's condition during the lessons/tests will be perhaps the best alternative to the observation. The student will not even notice that his or her condition is being analysed in order to further provide the collected information to the teacher and the PMPI for analysis and processing, according to which the curriculum will be adjusted individually for each student [16]. Instead, the system will teach the child to correctly recognize the emotional state of others and to respond appropriately to it. The main sections of the lessons will be the "Introduction", which includes introducing yourself and exploring yourself (self-identification), the "Main Part", in which the child will become acquainted with feelings and emotions, and the "Conclusion" in which the child can learn to distinguish friendship and love [17].

The main functionality includes [18]:

- Introducing children to emotions: anger, happiness, disgust, sadness, fear, surprise.
- Enriching the emotional sphere, improving the emotional state of children.
- Responding to negative emotions that interfere with the proper personal growth of the child.
- Reducing anxiety, overcoming fears.
- Learning to recognize the emotional manifestations of other people on various grounds (facial expressions, voice, body movements, etc.).

For a comprehensive picture of the essence of the system under study, we depict a tree of goals (Fig. 1)

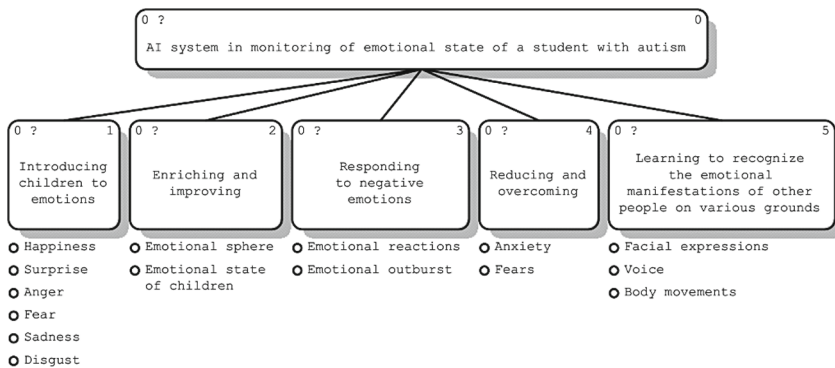


Fig. 1. Goal tree

There is a wide variety of methods for creating process diagrams. It was decided to use a data flow diagram (DFD) to describe this system, as it would be most convenient to depict the processes, relationships, and interactions between them.

The entity of “Student with autism”- in this case, is a child with ASD who needs help in expressing emotions to facilitate the learning process.

The entity of Teacher is important for controlling the learning process. It receives definitive recommendations from the system for further work and the direction in which it is necessary to promote further learning.

The entity of “PMPI” is a psychological, medical and pedagogical institution, that is, an organization comprising specialists of medical, pedagogical and psychological profile. The main task of the commission is to identify the causes of problems in the child’s education and to recommend him an educational program that he can successfully master (Fig. 2).

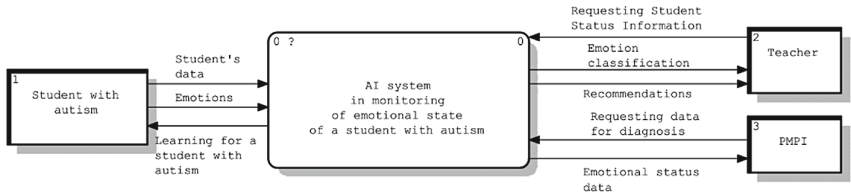


Fig. 2. Contextual diagram of data flows

In order to describe the next level of DFD, you have to decompose the main process into five others. These will include the following processes: “Download video with emotions”, “Set options for learning neural network”, “Analyse images using a neural network”, “Get emotional statistics”, “Provide guidance on learning through AI”. For reliable implementation, we also have to add two databases: “Emotions”, “Data from students with autism” (Fig. 3).

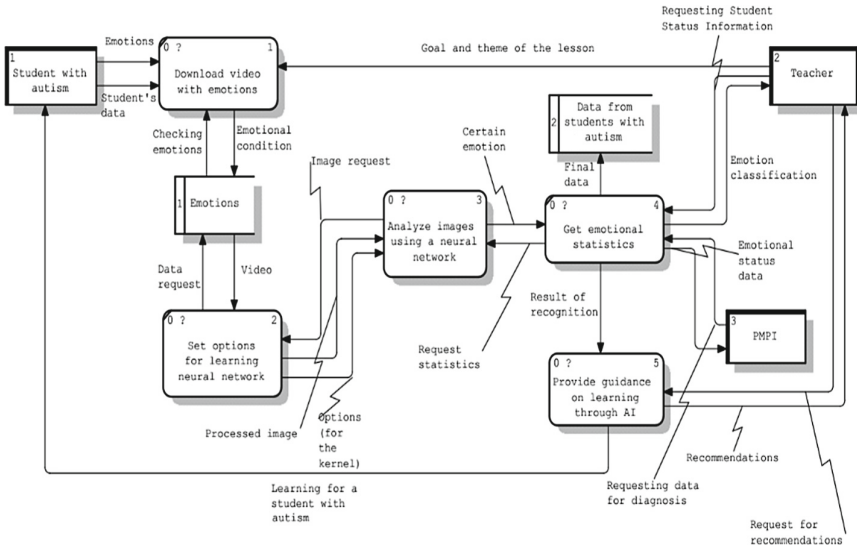


Fig. 3. DFD1

Sometimes, to fully describe the functioning of the system, decomposition must be performed more than once. Usually, this process is not repeated more than eight times, as it can lead to overloading of information in system and, as a result, will increase the complexity of reading the chart. Let’s perform a decomposition operation for the process “Analyse images using a neural network”. The result is a new five subprocesses: “Overlay the kernel over the input image”, “Compute the product of each number in the kernel and the image”, “Calculate the singular by summing the products”, “Set the number to the convolutional output”, “Go to the next section of the image” (Fig. 4).

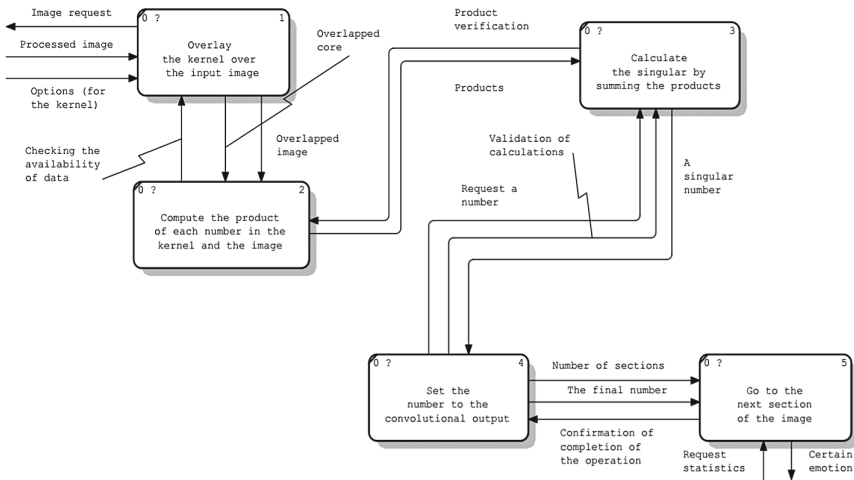


Fig. 4. Detailing the process of “Analyse images using a neural network”

For the clarity, it is also necessary to decompose the fifth process, “Provide guidance on learning through AI”. As a result, we get four subprocesses that will neatly describe its functionality. These include: “Obtaining the result of recognizing the emotions of a student with autism”, “Analyse the key moments of a student’s emotional reactions”, “Prepare a personal teaching methodology based on a recognized emotional state”, “Make recommendations to improve the response of students with autism” (Fig. 5).

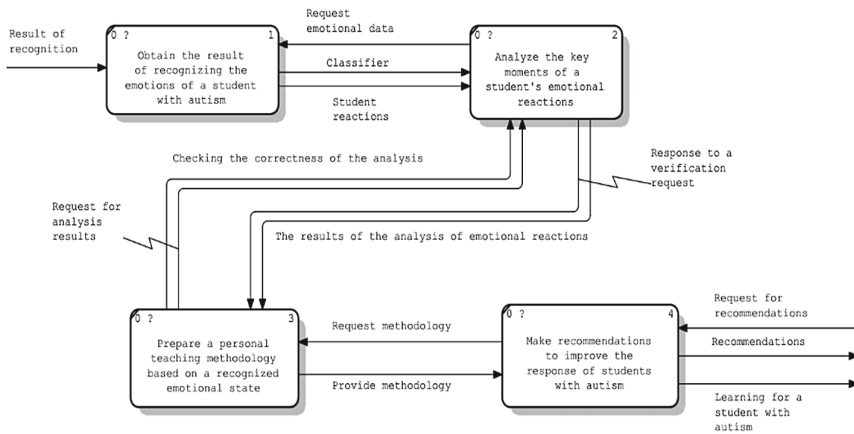


Fig. 5. Detailing the process of “Provide guidance on learning through AI”

As a result, the DFD diagram describes the basic functionality of the system. DFD allows to understand how an information system works, because it describes data flows and processes. It accelerates all subsequent stages of implementation. That is why the

initial illustration of the basic functional elements of the system is one of the most important stages of product development [19, 20] (Fig. 6).

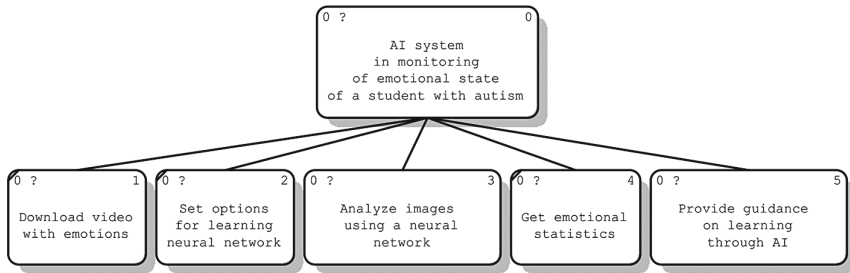


Fig. 6. The hierarchy of tasks of this system

3 Formulation of the System Design Process

Some sources are investigating the fact that in most systems, the mechanism for creating facial features is generated manually, which is a difficult task as this part is crucial in the further processing of data and the definition of emotions. In [29], the facial features are automatically extracted thanks to in-depth learning of the convolutional neural network. The method of using a convolutional neural network is a multilayer architecture. Each layer has a purpose. The purpose of convolutional layers is to extract facial features and image patterns [24]. Combined layers are created to reduce the number of resulting features. The neural network layers are designed to classify data using the data from previous layers' results. The system will then compare the images received with those in the database and record the new emotion [21–23].

The convolutional layer performs a mathematical operation composed of 3 elements: the first is an input, which is habitually expressed by a multidimensional array of data. And the second, it is the core, which is a multidimensional array of parameters accepted by the learning algorithm. The final element is a feature map. Multidimensional arrays are called tensors. The idea behind the network is that the core can recognize the visual patterns that come from the input (edges, lines, colors, etc.) and, therefore, be able to differentiate visual patterns between various forms of objects [44]. The process looks like this: first, the kernel laps the source image. Secondly, the product of each number in the kernel and each number in the overlapped input image is calculated. And thirdly, a unique number is calculated by summation these products. Last but one the resulting number is set to convolutional output. Finally, the kernel goes to the next section of the input image [45].

In Fig. 7 is a diagram of the use cases for the system under development. The cast includes: “Student with autism”, “Teacher”, “PMPI”.

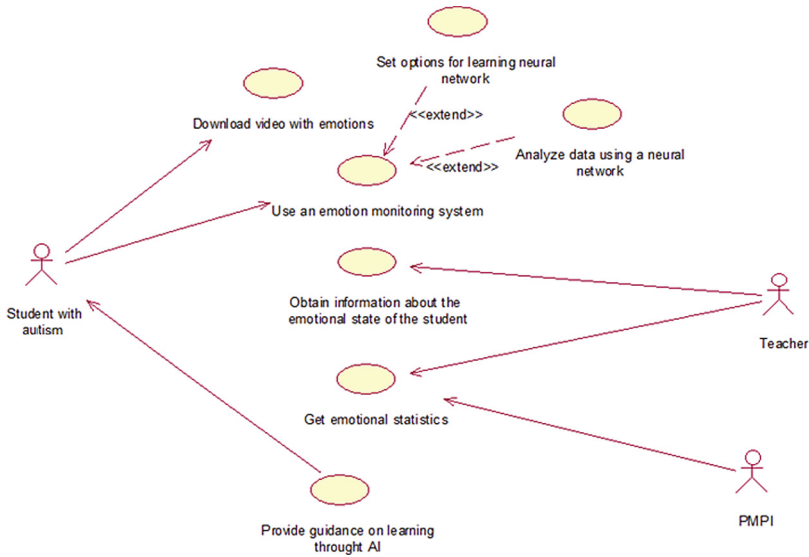


Fig. 7. Use Case

The Face Expression Database. There are several databases of possible facial expressions. Few of them include faces with different emotions (surprise, interest, joy, etc.), others have so-called spontaneous emotions, where reactions to certain situations are represented on the face [18]. Cohn-Kanade (CK) and CK plus (CK +) are databases that contain six basic emotions and already have action units (AU) [41]. The images are presented in sequence in the FACS. Each expression first appears as neutral, then switches to an expression of more intense emotion. The CK + version added spontaneous expressions recorded by 84 new people while being distracted during each photo session. Furthermore, the Radboud Face database (RaFD) provides photos of the eight primary emotions. The photos were taken with the help of locals, adults and children from the Netherlands. Participants demonstrated facial mimicry with three viewing directions and five camera angles. CMU Posed, Illumination, and Expression (PIE) database used synchronized multi-camera system (13 cameras) to obtain a large variety of poses (63 participants). Additionally, a flash system was used to obtain 21 illumination conditions for two background lightings. The talking variation was videotaped using 3 camera perspectives [30, 34, 35] (Fig. 8).

Since there is no way to create own emotion image database, the most popular and complete database is selected for implementation.

Since CK + contains the six basic emotions and spontaneous facial expressions, this database of emotion images is best suited for this project.

Facial Expression Recognition Algorithm. Facial expression analysis is usually frame-based. However, changing emotions themselves is a dynamic procedure. To recognize the current facial expression, a set of images should be selected for a short period. For the correct evaluation of emotions, it is convenient to take animation blocks and three-dimensional dots from the face for the last 30 frames [25–27].