INFORMATION SYSTEM FOR AUTOMATED CREATION AND ANALYSIS OF SOCIAL NETWORK MESSAGES CORPUS

UDC 004.9+81'27+81'33+81'42

DOI: https://doi.org/10.35546/2313-0687.2019.25.48-57

Borysova Natalia
PhD, associate professor of the Department of Intelligent Computer Systems, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine, **E-mail:** borysova.n.v@gmail.com, ORCID: https://orcid.org/0000-0002-8834-2536

Melnyk Karina
PhD, associate professor of the Department of Software Engineering and Management Information Technologies, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine, **E-mail:** karina.v.melnyk@gmail.com, ORCID: https://orcid.org/0000-0001-9642-5414

Melnyk Viktoriia
principal of Kharkiv general education school № 145, Kharkiv, Ukraine, **E-mail:** v13121423@gmail.com

Abstract. The purpose of work is development of information system for automated creation and analysis of corpus of social network messages. The information system allows getting statistical information about slang words and expressions used by youth. Combinations of multiple methods of corpus linguistics and sociolinguistics (questionnaire, associative experiment) have been used to achieve this goal. The core achievement of the work is the information system developed to resolve tasks of corpus creation and analysis. Texts for the corpus are collected by the information system from user-selected sources according to user-selected criteria. The information system can analyze the whole corpus or several parts of corpus according to the user-selected criteria. Corpus analysis results can be used by linguists-analysts, specialists in the field of automated natural language processing, corpus linguistics, sociolinguistics, as well as other interested experts and specialists to observe the development of any sociolinguistic phenomena. The scientific novelty of the research results consists in improving of the linguistic corpus automated formation and analysis technology, which allows increasing the information processing speed. The practical importance of the research results consists in the formation of algorithmic, linguistic, information support and software of information system for corpus automated creation and analysis.

Keywords: *automated natural language processing, corpus linguistics, sociolinguistics, information system, texts corpus, world using statistical characteristics, social networks messages analysis, youth slang.*

Introduction. The use of modern information technology today allows expanding and deepening scientific research in all scientific area. For linguistic research information technologies offer their technical capabilities for processing, storing and selecting language material [2; 3]. In addition, the involvement of information technology

allows obtaining objective conclusions about the language units functioning, helps to formulate qualitatively new conclusions about language and outlines new directions of language material research. This is made possible also by carrying out corpus researches that allow to abstract from the researcher's subjectivity and to approach the objective study of the language, since they rely mainly on real «living» language material, rather than on language intuition and introspection. Corpus researches imply the creation and use of linguistic corpuses – electronic texts collections selected by a specific criterion, and tagged depending on the research purpose. In our work, information technologies and linguistic corpus are used to study such a sociolinguistic phenomenon as youth slang. The relevance of the research topic is determined by the processes taking place in modern society. The democratization of society has increased the role of spoken language and its wider use in all areas of language communication. As a rule, the modern living language is formed by the most active representatives of the language society – journalists, politicians, businessmen and, first of all, young people. The language of youth, after the language of the media, is the second most influential on the state of modern living language of the mass communication sphere. Nowadays, non-literary words appear more often on media pages, in the influential people speeches and so on. Although research on youth slang has received some attention, this sociolinguistic phenomenon is still remains poorly understood. There was also a need to involve regional material more actively in researching this issue, as the specificity of the linguistic and cultural situation is generally determined by the regional context.

Last researches and publications analysis. The analysis of publications and researches [5-7] showed that to study such sociolinguistic phenomenon as youth slang, researchers use different approaches: the first group includes traditional sociolinguistic (questioning, associative experiment, surveys, field studies, etc.), and the second – modern approaches of corpus linguistics, which provide sufficient opportunities to explore different sociolinguistic phenomena at a qualitatively new level. However, despite the obvious prospects and effectiveness of using corpuses to study sociolinguistic phenomena, particularly in detecting language variations and changes, analyzing of language phenomena using by the person, social group, nation, etc., there are practically no similar works for Ukrainian language. There is no information on corpuses for sociolinguistic researches, on information systems for analyzing such corpuses, etc. Regarding to youth slang, it became only an object of lexicographic description, its lexical composition, formation and functioning were examined, but there is no information on the study of this phenomenon by corpus linguistics methods and approaches. Therefore, the results of last researches and publications analysis have determined the direction and purpose of our work.

The purpose of work is development of information system for automated creation and analysis of linguistic corpus.

The main material. The first stage of development process for resolving the task of automated creation and analysis corpus is creating of Software Requirement Specification (SRS). The SRS is a detailed description of the future information system with its functional and non-functional requirements. The list of requirements in formal view for the information system is presented in Fig. 1 in the form of a requirements diagram.

Let's consider this diagram in more detail way. The functional requirements involve the process of corpus automated creation and analysis. These possibilities are assigned with the requirements «Corpus automated creation» and «Corpus automated analysis» accordingly. The management of corpus' text is responsible for following functionalities: «Corpus' texts structure control» and «Corpus' texts management». These options allow using, adding, editing and deleting texts for corpus changing. A youth slang dictionary can be downloading. The functional requirement «Dictionary downloading» is responsible for this action.

Non-functional requirements define the information system quality attributes. They are important to ensure the usability and effectiveness of the entire information system. The reliability, performance, efficiency, usability and security are provided by appropriate options in information system.

Appropriate linguistic support is used for the information system correct work. It consists of our own social networks messages corpus, which was created by the information system, a youth slang dictionary and a slang database, which stores information about the inflection of words from youth slang dictionary.

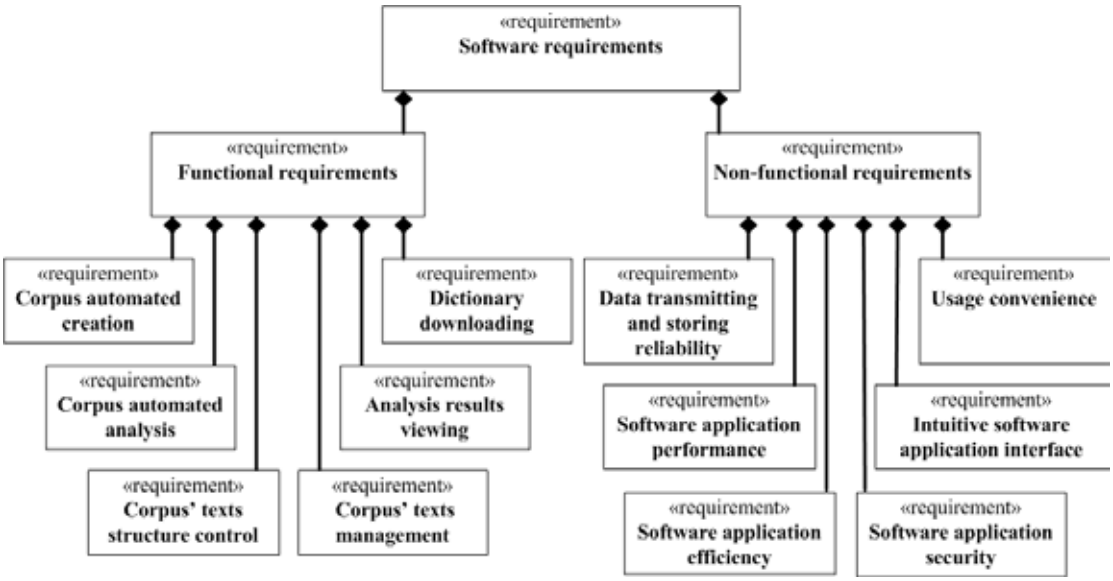

Fig. 1. Requirements diagram

As we research such a sociolinguistic phenomenon as youth slang as a texts sources for our own corpus have been selected social networks Facebook and Instagram messages and posts from various publics, such as «JustStory», «Убивчі Історії», «Веселі історії з життя людей», «Типовий Харків», «Типова Салтівка», «Підслухано ХПІ», etc., as well as the messages and posts from Pikabu web-site [1]. In order for the results of the study to be considered reliable when creating our own corpus, the general requirements for the corpuses were taken into account, namely:

- corpus volume – from 30000 to 1000000 words;
- the texts in the corpus must be representative;
- the size of the texts should be approximately the same;
- information can be extracted from the corpus.

Let's consider the process of automated corpus creation in more detail way. The analytical review of this process is presented using the IDEF0 diagram, which allows creating an appropriate functional model (Fig. 2).

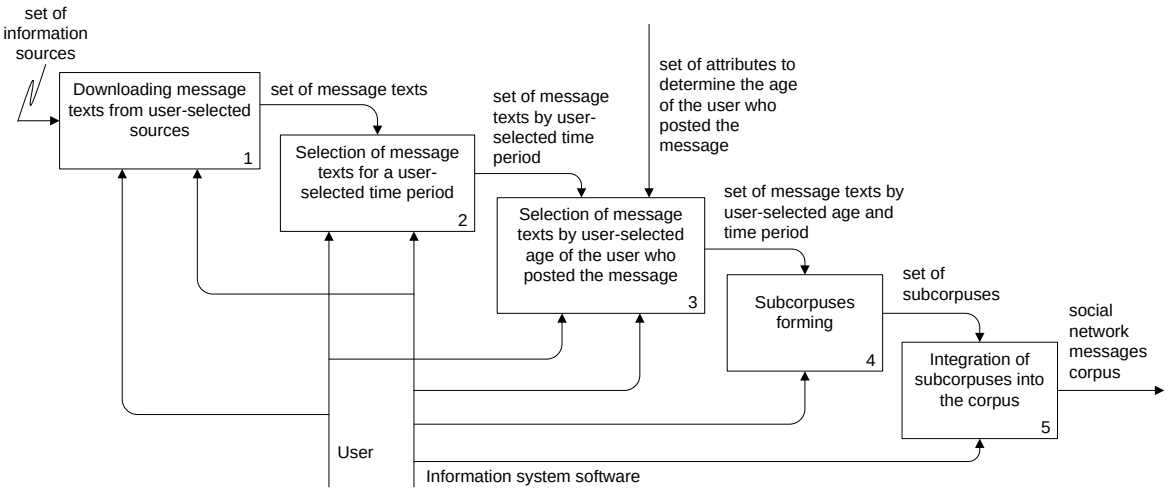

Fig. 2. The process of corpus automated creation

For corpus creation, first of all, user must specify a list of sources from which the texts will be selected. All found messages from the user-selected web pages are extracted and downloaded by the information system. Next, the user must specify for what time period he wants to select the messages (in our work – from December 2016 to December 2018 inclusively). All found messages from the user-selected time period are extracted by the information system. In the next step, the user must specify the age criterion that is what age of the people who posted the messages he wants to select the messages (in our work – from 16 to 25 years old inclusively). All found messages from the user-selected age are extracted by the information system. Then subcorpuses are automatically formed by information system. Texts from one source are stored in one subcorpus. In the last step, subcorpuses are integrated into the corpus.

Our own social networks messages corpus created by the information system fully complies with the general requirements for text corpuses. It contains approximately 150000 words, consists of subcorpuses, which stores texts from relevant sources, each subcorpus contains txt files, each file contains one comment or one message, the number of words in the file ranges from 150 to 300, the file size ranges from 2 to 3 Kb.

Another component of the information system linguistic support is the youth slang dictionary, which was obtained as a result of social networks users' survey, which was conducted from December 2016 to December 2018 inclusively. The questionnaire included a request to the respondent to fill in the questionnaire, questions about respondent age, place of residence, social status, and finally a request to write the 10 slang words, that he uses most often, to give a definition to these words and indicate their emotional tone (positive, negative, neutral). Since the survey results are stored in Google Tables, it is easy to use them later. Over a thousand people participated in the survey during the specified time period. The youth slang dictionary, which became a survey result, contains approximately five thousand slang words [2].

In order to justify that the youth slang dictionary can be used to analyze the created social networks messages corpus and the results of this analysis will be reliable, it is necessary to analyze the respondents who participated in the survey for dictionary creation. Google Forms has the suitable functionality for this. The respondents' analysis by age, social status and place of residence can be illustrated by the following charts (Fig. 3).

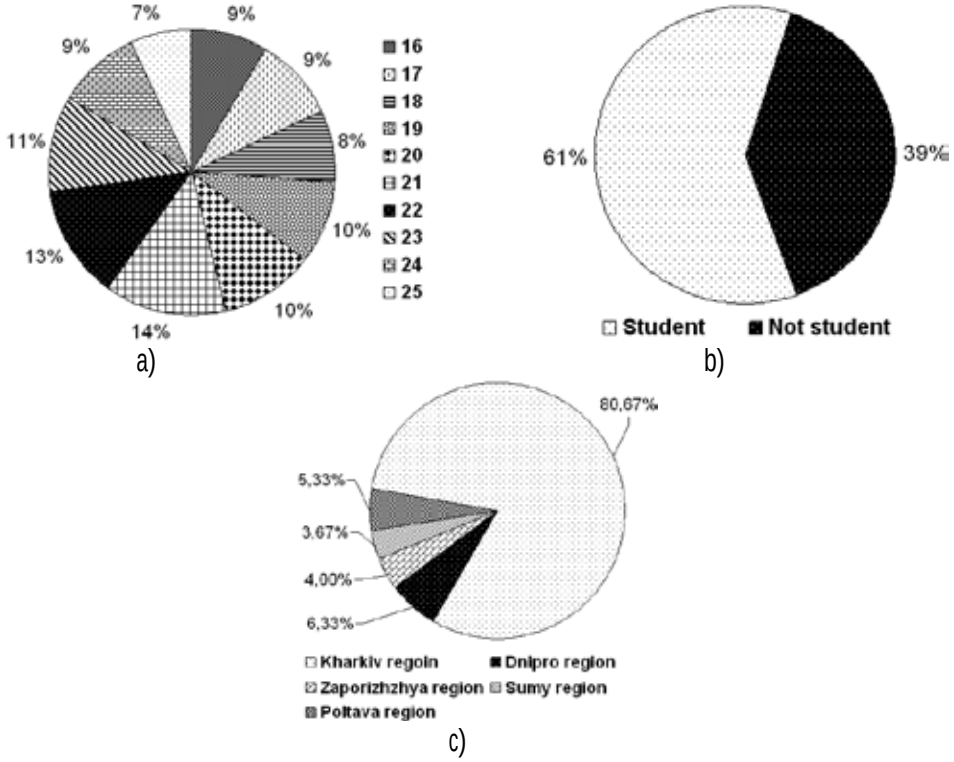

Fig. 3. The respondents' analysis: a) by age; b) by social status; c) by place of residence

Generalized analysis shows that a sample of respondents, which was formed in a random way, is representative. The respondents' analysis by age shows that respondents were separated into the approximately identical groups; there are no abnormally large or abnormally small groups. By social status almost two-thirds of the respondents are students, this fact has influenced on lexical composition of the dictionary by inserting slang words and expressions related to higher education. By place of residence almost 81% of respondents reside in Kharkiv region; this fact determined the choice of information sources for our own corpus.

In addition to our own corpus and youth slang dictionary, a slang database is also used as an information system linguistic support. Database stores information about the inflection of words from youth slang dictionary. The database structure is presented in Fig. 4.

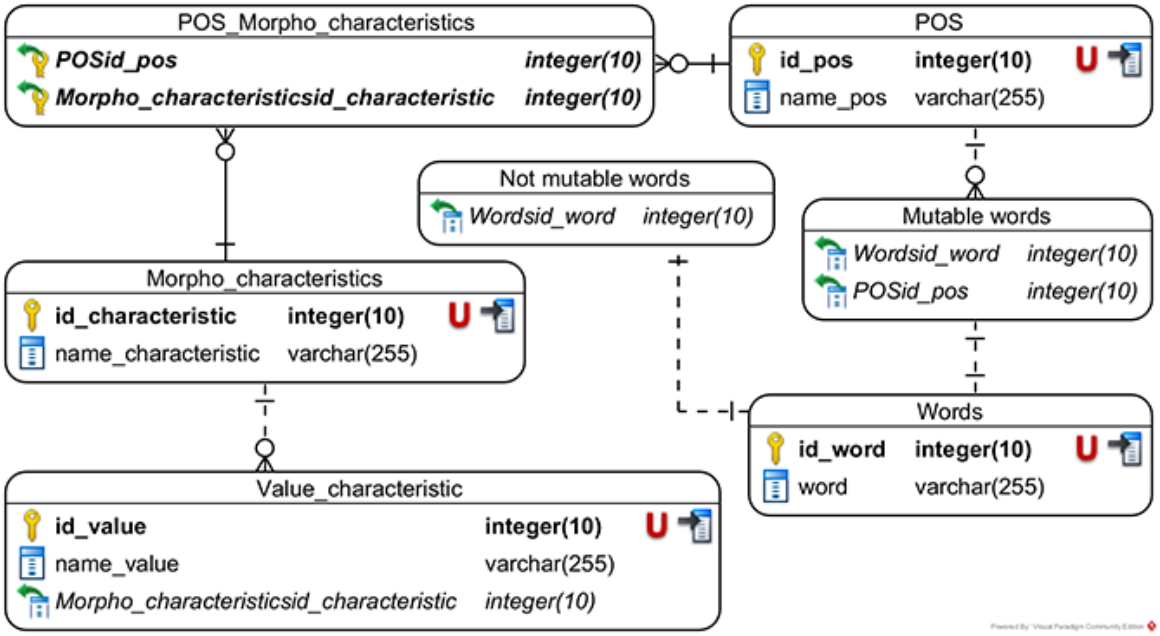

Fig. 4. The slang database structure

The slang database consists of seven entities, which are storing the following information:

– the entity "POS" represents all possible parts of speech;

– the entity "Morpho_characteristics" represents all possible morphological characteristics of parts of speech;

– the associative entity "POS_Morpho_characteristics" describes the relationship between the entities "POS" and "Morpho_characteristics";

– the entity "Value_characteristic" represents all possible values of each morphological characteristic;

– the entity "Words" is the categorical table, which describes mutable and not mutable words;

– the entities "Mutable_words" and "Not_mutable_words" are the child tables of the table "Words".

Detailed description of entities and their attributes are presented in Table 1.

Information about slang words inflection is needed because they may not necessarily appear in the text in their initial form, as in the dictionary. All the slang words were separated into two groups: the first group included mutable words which are submitted to regular rules of inflection for natural language parts of speech, and the second group included immutable words which are not submitted to these rules. For the mutable words the part of speech and morphological characteristics are indicated. The rules of inflection for each part of speech are programmed. Therefore when the information system analyzes the corpus, all forms of the mutable word are taken into account, but only one form is taken into account for the immutable word. The information system accesses the database when searching for slang words from the youth slang dictionary in corpus' texts.

Table 1

Description of client database model

Entity	Attribute	Description of attribute
POS	id_pos (PK)	Id of part of speech
	name_pos	Name of part of speech
Morpho_characteristics	id_characteristic (PK)	Id of morphological characteristic
	name_characteristic	The name of morphological characteristic
POS_Morpho_characteristics	id_pos (FK)	Id of part of speech
	id_characteristic (FK)	Id of morphological characteristic
Value_characteristic	id_value (PK)	Id of morphological characteristic value
	name_value	Name of value
	id_characteristic (FK)	Id of morphological characteristic
Words	id_word (PK)	Id of word
	word	Name of word
Mutable_words	id_word (FK)	Id of mutable word
	id_pos (FK)	Id of part of speech
Not_mutable_words	id_word (FK)	Id of immutable word

Considered linguistic support allows the information system to analyze the information from our own social networks messages corpus. The analytical review of corpus analysis process is presented using the IDEF0 diagram (Fig. 5).

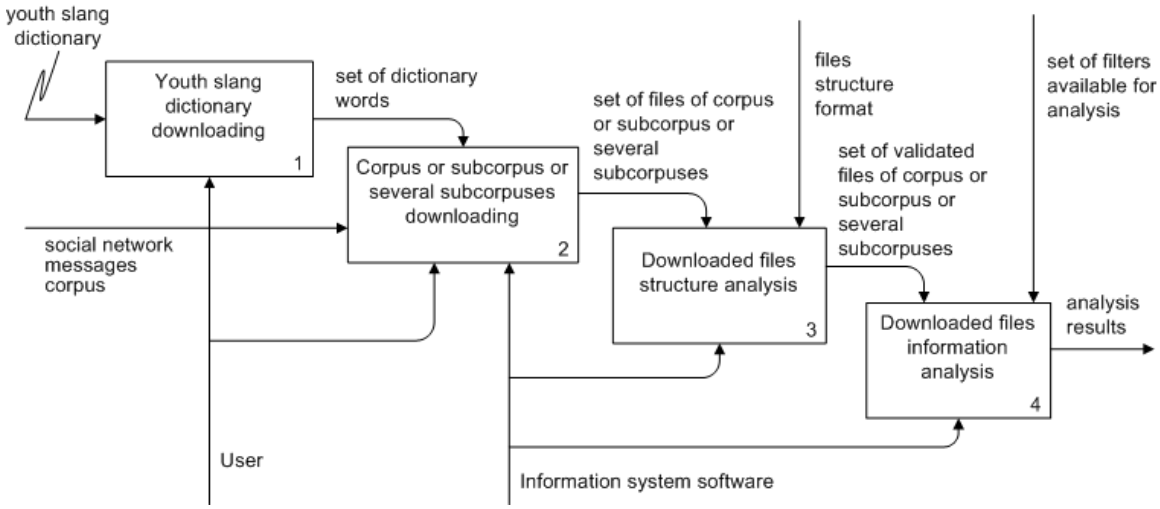

Fig. 5. The process of corpus automated analysis

In order to analyze the corpus, the user must first download a youth slang dictionary. After successful download, user can view the complete list of dictionary words, how many times the respondents have used the word in the questionnaire, the words' definitions (if there are several definitions for one word, they are separated by the symbol "|") and the emotional tone of the words. Also user can sort the list of words: alphabetically ascending or descending, as well as the number of using ascending or descending.

Then the user must download the corpus or subcorpus or several subcorpuses. After download, the user can view the list of downloaded files, can open or delete any number of files from this list. In the case, if the user has downloaded

files other than txt, empty files or files with the wrong structure, he will receive an error message. In addition, the user can view a list of such files, open or delete any number of files from this list.

In the next step, user can analyze the downloaded corpus or subcorpus or several subcorpuses. Various filters and any combination of them are available to the user: filters by age and time period. The user can request a complete list of words, TOP-10 words, or make an analysis by one word.

If available age filters are represented as an ordered set A with such elements A = (16, 17, 18, 19, 20, 21, 22, 23, 24, 25, all ages), the available time period filters are presented as an ordered set B with such elements B = (certain month, certain year, all period), and the available filters by word amount are represented as an ordered set C with such elements C = (one word, TOP-10 words, all words), then the Cartesian product of these three sets will represent all possible filter combinations available to the user if the user analyze the corpus by three filters. If two filters are used, all possible filter combinations available to the user will be represented as a Cartesian product of two sets: the set C (obligatorily) and the set A or the set B.

As a result of the analysis of the downloaded corpus, subcorpus or several subcorpuses, the user receive frequencies of use by youth in social networks messages of slang words and expressions from the youth slang dictionary.

For realization of the main tasks of the research the following architecture of the information system for automated creation and analysis of social networks messages corpus was chosen (Fig. 6).

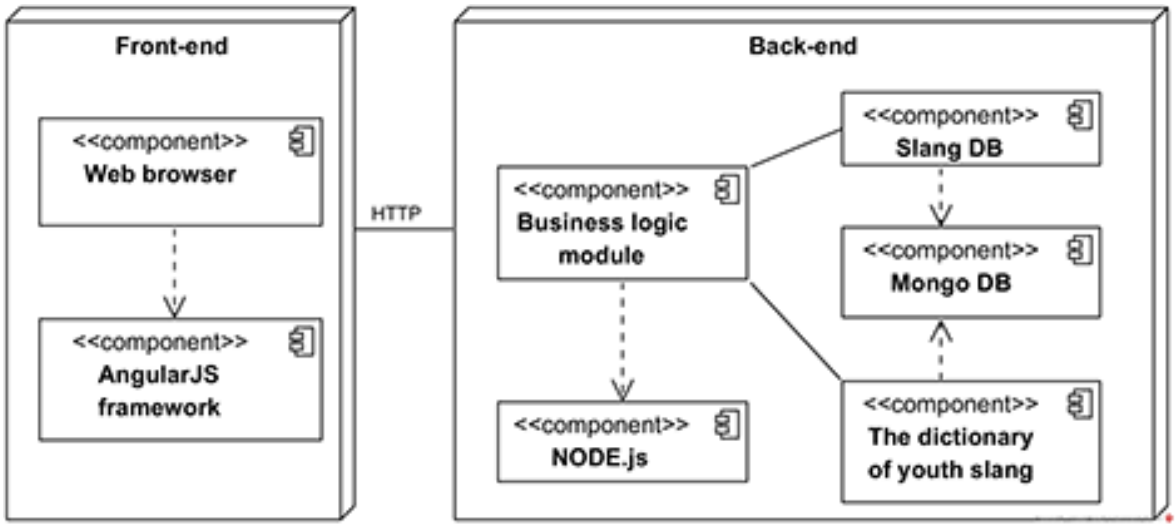

Fig. 6. Deployment diagram of information system

It is a web application, because information system performs specific functions by using a web browser as its client. The following technologies were suggested for information system implementing:

– HTML is using to represent the structural layer of the web application;

– CSS is using to implement the web application presentation layer;

– JavaScript is using to realize the behavior layer.

JavaScript and its frameworks can help to implement front-end and back-end of web application. AngularJS is a structural framework for dynamic web apps. It lets to create functionality with management corpus' data and the youth slang dictionary.

Back-end part of software uses JavaScript as well. Node.js is the server-side JavaScript runtime environment. It includes everything to execute the software written in JavaScript.

The domain area in our research is natural language processing. Therefore, database management system should work with complex structures of information with hierarchical relationships. So, MongoDB was proposed as a database management system. It uses for modern web applications.

Therefore, the considered linguistic, algorithmic, information support and software are enough to solve the main tasks of our research.

Let's consider the automated corpus analysis results gotten with information system in more detail way. In this paper only results of whole corpus analysis are presented.

A corpus analysis results by all age categories and all time period showed that the biggest use frequency had slang words which characterize emotions, assessments, expressing users' opinions, addressing to an interlocutor, as well as words related to computer using for games and work. Regarding the emotional tone of these words, they are mostly neutral words (57%), the number of positive and negative words is approximately the same (24% and 19% respectively).

To define if there is a difference in using slang words by young people of different age categories, the frequency words use by persons of age categories 16 and 25 years, that is, the youngest and oldest respondents, was determined. An analysis results showed that sixteen-year-old persons used slang words to express their emotions, attitudes, thoughts, and to address to another person. 40% words they use was a negative tone words, other 60% were divided equally between positive and neutral tone words. As for twenty-five-year-old persons, they used slang words related to computer using, to characterize their attitudes and emotions, to express their opinions, to address to an interlocutor. Most of them used words with a neutral tone (70%), least of all – with a positive (10%) and 20% of words with a negative tone. This difference can be explained by the peculiarities of the psychic and psychological development of these age categories people. Thus, the results obtained are completely correlated with well-known confirmed psychological theories of personality development.

Also we got interesting results of corpus analysis for different time periods. Let's consider the corpus analysis results by all age categories for July 2018 in more detail way. It was found that in this time period users used slang words related to computers and computer games, to express emotions, mostly positive, opinions and to address to another person. Regarding the emotional tones of these words, they used 60% of words with a positive tone, 30% with a neutral, and 10% with a negative one. Such results in lexical composition and emotional coloring of words can be explained by the fact that July is a vacation time for schoolchildren and students.

To define if there are changes in lexical composition of slang words using during one year, the corpus analysis results by all age categories were compared for 2017 and 2018. It was found that about 30% of slang words on the list for 2017 continued to be used in 2018. Regarding the emotional tones of the words, in 2017 it was used 50% of negative words, 40% of positive and 10% of neutral one. In 2018, the situation changed slightly and more words with a neutral tone (60%) began to be used, positive and negative words were used equally (20% each).

Conclusions. Therefore, the information system for automated creation and analysis of social networks messages corpus has all the necessary functionality for comprehensive corpus analysis, in particular it can be used to study the frequency characteristics of the specific lexis use by a certain group of native speakers. To do this, the system selects the texts for the corpus from the user-selected sources by certain criteria, and then analyzes the corpus information by user-selected parameters.

REFERENCES:

1. Борисова, Н.В., Ніфтілін, В.В. (2018). Застосування методів корпусної лінгвістики для дослідження особливостей використання сучасного молодіжного сленгу. *Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: тези доповідей XXVI міжнародної науково-практичної конференції MicroCAD-2018*, *Ч. І*, 27.
2. Борисова, Н.В., Ніфтілін, В.В. (2017). Автоматизоване створення електронного словника *Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: тези доповідей XXV Міжнародної науково-практичної конференції MicroCAD-2017*, *Ч. І*, 32.
3. Мельник, К.В. (2018). Метод інформаційного скринінгу медичної документації. Под ред. В.С. Пономаренко, *Інформаційні технології: сучасний стан та перспективи* (С. 391-406). Харків, Україна: ТОВ «ДІСА ПЛЮС».
4. Мельник, К.В. (2017). Моделювання процесу інтелектуальної обробки медичних даних. *Системи обробки інформації*, *4* (*150*), 237-244.
5. Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge, England: Cambridge University Press. doi:10.1017/9781316410899

6. Crawford, W., Csomay, E. (2015). *Doing Corpus Linguistics*. New York, NY: Routledge. doi: 10.4324/9781315775647
7. Friginal, E. (2017). *Studies in Corpus-Based Sociolinguistics*. New York, NY: Routledge. doi: 10.4324/9781315527819

ІНФОРМАЦІЙНА СИСТЕМА АВТОМАТИЗОВАНОГО СТВОРЕННЯ ТА АНАЛІЗУ КОРПУСУ ПОВІДОМЛЕНЬ СОЦІАЛЬНИХ МЕРЕЖ

Борисова Наталя Володимирівна
кандидат технічних наук, доцент кафедри інтелектуальних комп'ютерних систем,
Національний технічний університет «Харківський політехнічний інститут», м. Харків, Україна,
E-mail: borysova.n.v@gmail.com, ORCID: https://orcid.org/0000-0002-8834-2536

Мельник Каріна Володимирівна
кандидат технічних наук, доцент кафедри програмної інженерії та інформаційних технологій управління,
Національний технічний університет «Харківський політехнічний інститут», м. Харків, Україна,
E-mail: karina.v.melnyk@gmail.com, ORCID: https://orcid.org/0000-0001-9642-5414

Мельник Вікторія Іванівна
директор Харківської загальноосвітньої школи I-III ступенів №145, м. Харків, Україна,
E-mail: v13121423@gmail.com

Анотація. Метою статті є розробка інформаційної системи автоматизованого створення та аналізу корпусу повідомлень соціальних мереж для отримання статистичної інформації щодо використання молоддю сленгових слів та виразів. Для досягнення поставленої мети комплексно використовувались методи корпусної лінгвістики та соціолінгвістики (анкетування, асоціативний експеримент). Основним результатом дослідження є розроблена інформаційна система, яка виконує дві основні задачі: створення корпусу та аналіз корпусу. Тексти для корпусу збираються інформаційною системою з зазначених користувачем джерел за визначеними користувачем критеріями. Щодо аналізу інформації корпусу інформаційна система може аналізувати за обраними користувачем критеріями як весь корпус, так і його частину, або декілька частин. Результати аналізу корпусу можуть бути використані лінгвістами-аналітиками, спеціалістами в галузі автоматизованої обробки природної мови, корпусної лінгвістики, соціолінгвістики, а також іншими зацікавленими експертами та спеціалістами для моніторингу розвитку будь-яких соціолінгвістичних явищ. Наукова новизна одержаних результатів дослідження полягає у вдосконаленні технологій автоматизованого формування та аналізу лінгвістичних корпусів, що дозволяє підвищити швидкість обробки інформації. Практична значимість одержаних результатів полягає у формуванні алгоритмічного, лінгвістичного, інформаційного та програмного забезпечення інформаційної системи автоматизованого створення та аналізу корпусів текстів.

Ключові слова: *автоматизована обробка природної мови, корпусна лінгвістика, соціолінгвістика, інформаційна система, корпус текстів, статистичні характеристики використання слів, аналіз повідомлень соціальних мереж, молодіжний сленг.*

ИНФОРМАЦИОННАЯ СИСТЕМА АВТОМАТИЗИРОВАННОГО СОЗДАНИЯ И АНАЛИЗА КОРПУСА СООБЩЕНИЙ СОЦИАЛЬНЫХ СЕТЕЙ

Борисова Наталья Владимировна
кандидат технических наук, доцент кафедры интеллектуальных компьютерных систем,
Национальный технический университет «Харьковский политехнический институт», г. Харьков, Украина,
E-mail: borysova.n.v@gmail.com, ORCID: https://orcid.org/0000-0002-8834-2536

Мельник Карина Владимировна
кандидат технических наук, доцент кафедры программной инженерии и информационных технологий управления,
Национальный технический университет «Харьковский политехнический институт», г. Харьков, Украина,
E-mail: karina.v.melnyk@gmail.com, ORCID: https://orcid.org/0000-0001-9642-5414

Мельник Виктория Ивановна
директор Харьковской общеобразовательной школы I-III ступеней №145, г. Харьков, Украина,
E-mail: v13121423@gmail.com

Аннотация. Целью статьи является разработка информационной системы автоматизированного создания и анализа корпуса сообщений социальных сетей для получения статистической информации об использовании молодежью сленговых слов и выражений. Для достижения поставленной цели комплексно использовались методы корпусной лингвистики и социолингвистики (анкетирование, ассоциативный эксперимент). Основным результатом исследования является разработанная информационная система, выполняющая две основные задачи: создание корпуса и анализ корпуса. Тексты для корпуса собираются информационной системой из определенных пользователем источников по определенным пользователем критериям. Относительно анализа информации корпуса информационная система может анализировать по выбранным пользователем критериям, как весь корпус, так и его часть, или несколько частей. Результаты анализа корпуса могут использоваться лингвистами-аналитиками, специалистами в области автоматизированной обработки естественного языка, корпусной лингвистики, социолингвистики, а также другими заинтересованными экспертами и специалистами для мониторинга развития любых социолингвистических явлений. Научная новизна полученных результатов исследования состоит в усовершенствовании технологий автоматизированного формирования и анализа лингвистических корпусов, что позволяет повысить скорость обработки информации. Практическая значимость полученных результатов состоит в формировании алгоритмического, лингвистического, информационного и программного обеспечения информационной системы автоматизированного создания и анализа корпусов текстов.

Ключевые слова: *автоматизированная обработка естественного языка, корпусная лингвистика, социолингвистика, информационная система, корпус текстов, статистические характеристики использования слов, анализ сообщений социальных сетей, молодежный сленг.*

СПИСОК ЛІТЕРАТУРИ:

1. Борисова Н. В., Ніфтілін В. В. Застосування методів корпусної лінгвістики для дослідження особливостей використання сучасного молодіжного сленгу. *Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: тези доповідей XXVI міжнародної науково-практичної конференції MicroCAD-2018* / відп. ред. проф. Сокол Є.І. Харків: НТУ «ХПІ». Ч 1. С. 27
2. Борисова Н.В., Ніфтілін В.В. Автоматизоване створення електронного словника. *Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: тези доповідей XXV Міжнародної науково-практичної конференції MicroCAD-2017* / відп. ред. проф. Сокол Є.І. Харків: НТУ «ХПІ». Ч. І. С. 32
3. Мельник К. В. Метод інформаційного скринінгу медичної документації : монографія / відп. ред. В. С. Пономаренко. Харків : ТОВ «ДІСА ПЛЮС», 2018. С. 391-406.
4. Мельник К. В. Моделювання процесу інтелектуальної обробки медичних даних. *Системи обробки інформації*. 2017. № 4 (150). С. 237-244.
5. Crawford W., Csomay E. Doing Corpus Linguistics. New York: Routledge, 2015. 178 p.
6. Brezina V. Statistics in Corpus Linguistics: A Practical Guide. Cambridge: Cambridge University Press, 2018. 314 p.
7. Friginal E. Studies in Corpus-Based Sociolinguistics. New York: Routledge, 2017. 382 p.