Вычислительные технологии, 2022, том 27, № 2, с. 122–137. © ФИЦ ИВТ, 2022 ISSN 1560-7534
Computational Technologies, 2022, vol. 27, no. 2, pp. 122–137. © FRC ICT, 2022 eISSN 2313-691X

DOI:10.25743/ICT.2022.27.2.010

Conceptual design of the software system for automated complex analysis of poetic texts

Kozhemyakina O. Yu.
Federal Research Center for Information and Computational Technologies, 630090, Novosibirsk, Russia
Corresponding author: Kozhemyakina Olga Yu., e-mail: olgakozhemyakina@mail.ru
Received March 31, 2022, accepted April 7, 2022.

The paper presents the process of conceptual design and implementation of the software system for automated complex analysis of poetic texts. At the design stage, the tasks that the information system for presenting the results of complex analysis of poetic texts is designed to solve, as well as the requirements in order of priority for the software system. The developed system combines heterogeneous information about the results of the analysis of poetic texts obtained at each of the presentation levels. Based on the needs of potential users, the following tasks have been solved: the description of the external interacting elements of the system has been completed; the conceptual description of the system has been formulated; the interface for accessing the information system storage has been developed; the graphical interface of the information system has been implemented. The results of the work showed the fundamental possibility of integrating the system components. The convenient tool for philologists to automate the complex analysis of poetic texts in Russian has been created.

Keywords: natural language processing, software system design, text analysis information system.

Citation: Kozhemyakina O.Yu. Conceptual design of the software system for automated complex analysis of poetic texts. Computational Technologies. 2022; 27(2):122–137. DOI:10.25743/ICT.2022.27.2.010.

Introduction

The modern approach to the study of text messages involves the usage of a multilevel information model, one of the variants of which is described in the work of the German researcher W. Gitt [1]. The structure of the model is shown in Fig. 1.

As noted in [1], the lowest level of the presented model corresponds to the definition of C. Shannon's information [2], then, in ascending order, there are levels of syntactic, semantic and pragmatic data. For automated text analysis, the upper level — apobetics — is not considered, since either it coincides with the previous one, i. e. the action corresponds to the goal, or the goal is explicitly spelled out, or the message is not enough to recognize the goal and the additional data sources are needed to be used. It is worth noting that in order to have high-level information in a message, it is necessary, but not enough, to have the information from all previous levels, since its volume also depends on meta-information,

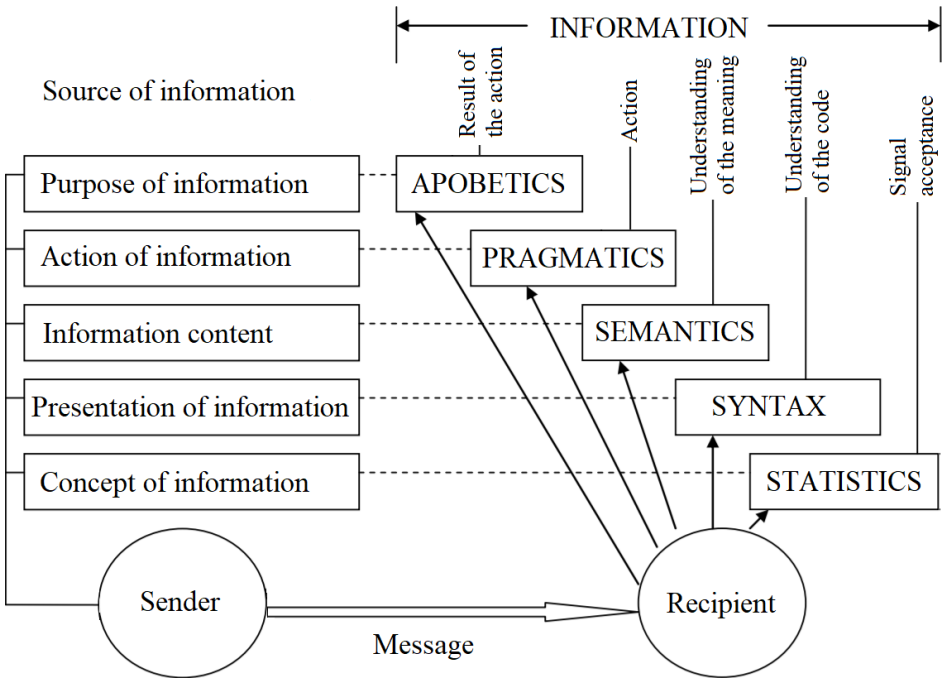

Fig. 1. W. Gitt's five-level information model

including the characteristics of the recipient, context, etc., what is true for all levels of the model. The two lower levels of the message (statistics and syntax) are directly related to the language (here language is meant as a sign system; it should be distinguished from subsets of the language that are called similarly, for example: spoken English, literary Russian, Pushkin language [3], etc.) and the encoding of the message. Their influence on higher levels is questionable, with the exception of literary texts, primarily poetic ones: in particular, Yu.M. Lotman argued that "the phenomenon of structure in verse always ultimately turns out to be a phenomenon of meaning" [4]. For a message, for example, of the scientific style there is no such influence: the same article can be translated into another language with minor changes, while practically losing its content, especially when it comes to exact sciences. In poetic texts, the levels of the structure of some message can be mapped into the levels of the structure of the verse, which also have the stable hierarchy [5]: metric, rhythm, phonetics, vocabulary, grammar, literary style, thematics, literary genre, while the levels of the structure of message and verse can be compared as follows: phonetics belongs to the statistical level, metrics and rhythmics — to the syntactic level (what is quite correlated with Fig. 1 (see above) from [1]), vocabulary and grammar — to the semantic level, on the border of the semantic and pragmatic levels is the thematics (see Fig. 2).

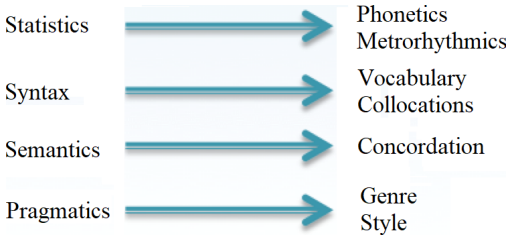

Fig. 2. The correlation of the levels of the poetic text to the levels of the scheme of W. Gitt

The process of the analysis of each poetic text consists of considering the levels as independent semantic units, followed by linking the data obtained with other elements of the structure. The vocabulary forms the semantics of this particular poem, the metric is the general background of the semantic tradition on which it is perceived. Thus, the study of the influence of the lower levels of the poetic structure on the higher ones is the actual problem of Russian philology. The beginning of a systematic study of the influence of the lower levels of the verse structure on the higher ones, despite the presence of separate studies, can be considered the work of K.F. Taranovsky [6]. In his speech at the Fifth Congress of Slavists in 1963, K.F. Taranovsky presented a report "On the interaction of poetic rhythm and themes", in which, based on the analysis of several Russian poetic texts, the interaction of rhythmic features and genre application of the pentameter trochee was investigated. It has been shown that in many poems written by this size (starting with "I go out alone on the road ..." by M.Yu. Lermontov), "the dynamic motive of the path is opposed to the static motive of life" (see [6]). In this work, the methodology was proposed for determining the semantics of a particular poetic size, which consists in studying not its single usage, but the tradition of its genre and thematic usage [5], what involves the analysis of the corpus of poetic texts.

One of the main difficulties encountered in solving this problem is the need to analyze the large corpuses of poetic texts. This task is extremely time-consuming, therefore, as a rule, only a relatively small circle of poems by poets of the classical period falls into the field of view of the researcher, what significantly reduces the completeness of the analyzed material and, consequently, the reliability of obtained results. Thus, there is a problem of automating the analysis of various levels of the verse structure, what will free the researcher from routine work and at the same time allow expanding the range of analyzed works. The above correlation between the levels of the structure of a message and a poem shows that many technologies and mathematical methods used in computer science can be also used in the process of automating the analysis of poems. Of course, the simplest mathematical approaches have long been used in the philological analysis of Russian poems. The frequency dictionaries of the language of classical poets are widely known. The numerous studies of statistics of types of Russian rhyme (including in relation to historical changes) are summarized in [7].

However, the statistical information is still collected, in most cases, manually. Some studies describing the integrated approach to automating the characteristics of Russian poetic texts (see, for example, [8]) usually affect very specific genres of poetry, such as folklore, whose structural characteristics, in particular, metrics, etc., differ significantly from the corresponding structures of literary verse. It should be noted that the research of foreign authors in this area in relation to the Russian language is unknown to us. The lexical analysis of the poem provides [9] for the creation of its lexical dictionary, which is used, in particular, to identify the dominant parts of speech, thematic (semantic) fields and poetic phraseology (primarily metaphors). Among the non-commercial software products that solve the task of compiling a lexical dictionary of a text, we can name the Yandex stemmer [10]. It allows to extract both words that are a given part of speech (what automatically solves the problem of identifying the dominant parts of speech) and the collocations of a given structure (for example, (adjective) + (noun) or (noun) + (noun in the genitive case)). Also, the last of these functions can significantly enrich the traditional dictionaries of the poet's language by supplementing them with collocations.

As for the tasks of identifying thematic fields and metaphors, their solution requires, along with lexical dictionaries of words and collocations, additional and often poorly formalized

information (for example, about the belonging of lexemes to a certain thematic field, semantic archetype, etc.), as a result of what we refer these tasks to the perspective of research. The grammatical analysis of the text includes the determination of the possible belonging of a text fragment to a nominal or verbal type (respectively, to continuous nominal sentences or enumeration of actions), as well as the time plan and subject structure of the poem (what requires the studying the usage of the categories of tense, voice and person), nominal and verbal types, in turn, are determined by direct analysis lexical dictionary, or morphological analysis. To determine the usage of the categories of time, pledge and person, it is necessary to additionally use the simple morphological rules of the Russian language that allow to establish the specific categories. The direct definition of the subject of a poem is a very difficult task for an automated solution, since it requires semantic analysis of texts at a level close to human perception of texts in natural language. However, the study of the dependence of the thematics on the lower levels of the poetic structure is one of the least studied areas of philological analysis. The usage of methods of statistical analysis of poetic texts of large volumes can also be an effective method of solving these and similar problems. The important area of research is the usage of multidimensional analysis of semantic, emotional and other characteristics, the large-scale application of which is almost impossible without the usage of the methods of automation. Working with large text corpora involves the usage of data mining methods. The modern approaches to clustering the text documents using several similarity scales are presented, for example, in the monograph [11].

1. Software system for presenting the results of automated complex analysis of poetic texts

Let's form the requirements and describe the design and implementation of the user interface of the system for presenting the results of automated complex analysis of poetic texts. First of all, let's clarify the definition of the term "information system". According to [11], the intelligent information systems not only extract information from data, but also gain new knowledge: the semantic information is transformed into data, but the reverse process is important for the user — "extracting information and knowledge from data that the user needs" [11]. In the context of our studies, we will use the term "information system" as the designation of the corresponding component of the software system for automated complex analysis of poetic texts.

Each module of the software system for automated analysis of poetic texts belongs to one of the structural levels of text analysis: structural, semantic, pragmatic. The structural analysis of the poetic text is associated with the extraction of its metrorhythmic characteristics. Within the framework of semantic analysis, there are studies on the separation of semantic constructions from a work; the pragmatic level includes studies on the automatic determination of high-level characteristics of a poetic text, such as genre and style. The process of designing and implementing of the information system for presenting the results of the analysis of poetic texts includes the formulation of tasks that the information system is designed to solve, as well as the presentation of requirements in order of priority for the overall project. The developed information system should combine heterogeneous information about the results of the analysis of poetic texts obtained at each of the presentation levels, according to the conceptual description of the system. The full implementation of the information system will provide a significant simplification of the research of poetic texts.

2. Overview of existing information systems

Let's mark the systems and studies in which the tasks of applying the methods of machine learning, approaches and designed systems were set, however, we emphasize that our system is of a pioneering nature due to its complex architecture, module structure and filling with algorithms that are most effective in analyzing the poetic text.

It should be mentioned, first of all, the article [12], in which a large program of research on metric, rhythmic and phonetic (including rhyme) characteristics of Russian poetic texts was outlined. This program, in turn, relied on the STARLING system [13], which was part of the project "Automated linguistic and poetic analysis of Russian poetic texts" (after the death of the project manager S.A. Starostin, the work in this direction was discontinued). It is on the basis of these studies that we have implemented the currently operating software tool for analyzing the metrorhythmic characteristics of poetic texts [14], described in [15]. However, the algorithm from [12] is semi-empirical, what reduces its accuracy in cases of complex accentuation, therefore, for further research, we implement a more strictly justified algorithm from the article [16], modified taking into account the ambiguous accentuation of texts in Russian.

A team of authors led by I.A. Pilshchikov and A.S. Starostin has achieved great success: since 2008, a number of works have been carried out on automatic meter recognition in syllabotonic verses [17–19]. In 2016, in the report "Instrumental environment for working with Russian-language poetic corpora and their specialized markup" [20], the instrumental computer environment "Workplace of a poet" was demonstrated; the possibility of heuristic accentuation of non-dictionary words, the creation of the template editor interface, the visualization of the results of automatic analysis of metrics and rhythmic verse is described. However, we are not aware of the works in which the authors conducted a study of the automation of the analysis of characteristics of a higher level (for example, the definition of genre).

The tasks of the analysis phonetic and lexical characteristics of a text are less specific and are solved much more often than the task of analyzing metrorhythmic characteristics characteristic of poetic texts, so we use more or less standard algorithms described in the article [15] to solve these problems. To determine the styles and genre characteristics of texts, we use the most well-known techniques of ensembling basic algorithms in composition, such as weighted voting, boosting and stacking [21], the similar approaches are used for other texts in [22, 23].

Let's consider, as part of the design task, some of the existing information systems that are designed to study certain characteristics of poetic texts: the project "Concordance to Lomonosov's texts", the SPARSAR system, the Metricalizer web application.

The project "Lomonosov Concordance" [24] was launched in 2009 and is based on a corpus of author's texts equipped with structural, philological and grammatical markup. This project is available via the web interface [25] and is both the alphabetical-frequency concordance to the texts and the collection of editions of Lomonosov's texts. The technological chain of work with the corpus includes a considerable part of manual markup of the corpus and text segmentation; morphological analysis is performed with the help of the parser, followed by post-processing (removal of homonymy, correction of errors). The practical implementation simplified the interaction with the created concordance — the user can interactively choose the appropriate term to work with it. However, the project was not completed by the authors.

The SPARSAR system described in R. Delmonte's work [26] assumes an automatic complex analysis of poetic texts in order to study their style. SPARSAR [27] analyzes each poem at different levels: at the sentence level, at the row level and at the stanza level. Such information system would be useful to the authors of the article, but the detailed description of the internal structure is given to the module associated with automatic text reading (TextToSpeech [28]).

The Metricalizer web application [29], developed by K. Bobbenhausen and B. Hammerich, allows to perform the automatic analysis of the metric characteristics of German verses [30]. The system provides the metrical analysis of a poem, the creation of XML documents based on the results of the analysis, the analysis of texts by accentuation and rhyming, the phonetic analysis of words in the X-SAMPA format (Extended Speech Assessment Methods Phonetic Alphabet). This system is most relevant to the structure that could be applied to our information system, however, the detailed structure of the Metricalizer system is not presented in [30].

The variety of algorithms for processing poetic texts is undoubtedly the important factor in the successful solution of the described problem. However, in many cases, the implementation of the algorithm in the form of the separated software product and its usage without the direct participation of the authors remains relevant. Because of this, the software application should be a kind of information subsystem with specific input and output data. It is important to configure the subsystem's interaction with both the end user and other subsystems.

3. Approach to the design and implementation of the information system

So, the approach to the design and implementation of the information system for the analysis of poetic texts should take into account its complex modular structure. Next, we will describe the process of designing the information system for presenting the results of automation of complex analysis of poetic texts.

The process of the analysis of poetic texts is reduced to the following sequential steps (see Fig. 3), on which the analysis of the characteristics of the text is carried out:

- the initialization — the formation of a corpus of poetic texts and its preprocessing for subsequent analysis;
- the structural analysis — the determination of low-level characteristics of a poetic text (phonetics and metrorhythmics of a poem);
- the semantic analysis — the definition of semantic constructions taking into account poetic synonymy;
- the pragmatic analysis — the expert assessment of belonging to certain stylistic characteristics for a poetic text (genre, style, etc.);
- the synthesis of the conducted research — the determination of the influence of lower levels of poetic texts on higher ones, as well as combining the results in a form convenient for perception and search.

The conceptual design outlined by us in [31] included the formation of the capabilities of the information system which should have the following features.

1. The providing of the access to the corpus of poetic texts. At the same time, the additional requirements may be imposed on the texts. For example, during processing,

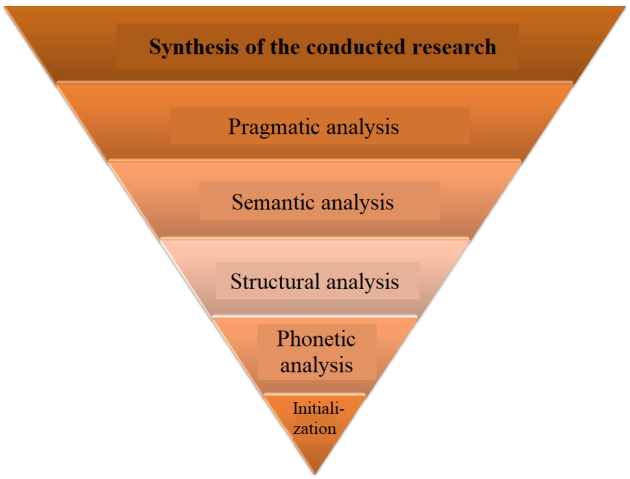

Fig. 3. Interrelations in the process of analyzing poetic texts

it is important to take into account the features of the old (of time of text creation) spelling.

2. The automated processing of the corpus of poetic texts stored in the database:
   a) the determination of the phonetic characteristics of the text;
   b) the research of metrorhythmic characteristics: metric, number of feet, rhyming of stanzas, etc. with indication of the ambiguities that cannot be resolved automatically;
   c) the determination of lexical characteristics of the text;
   d) the definition of genre and style characteristics of the text.
3. The entering the received characteristics into the storage (database).
4. The statistical processing of the received characteristics and their presentation in a form convenient for the researcher.
5. The ability to import poetic text corpuses from databases and files, and to export them to other information systems for further automated processing.

The algorithms implemented during the stages of analysis (structural, semantic, pragmatic) are described in our works [32, 33]. However, the task of the synthesis of studies within the information system suitable for usage by philological experts is a priority. When designing a system, we consider the project subsystems in the form of a "black box" — the leading role is played by the data supplied to the input and received at the output. The mind map and the use case diagram were used as the basic design tool at the initial stage. The process of the analysis of poetic texts is presented in the IDEF0 notation [34]. The following top-level business processes are highlighted.

1. The text preprocessing: the raw text (possibly in pre-reform spelling) with information about this text is received at the input. Under the control of the rules for the formation of pre-reform orthography and modern orthography, the text is converted into modern orthography with the results recorded in the storage system. Further text processing is carried out in modern orthography.
2. The structural analysis: based on the formalized rules for constructing the meter and rhythm of a poetic text, its structural characteristics are extracted: the type of rhyme, the number of female and male endings, etc.

3. The semantic analysis: on the basis of formalized rules for constructing phrases, the syntactic constructions are extracted from the poetic text and identified using the basic dictionary that takes into account synonymy.
4. The pragmatic analysis: using the classification system of genres and styles, a hypothesis is formed at the output about the belonging of a poetic text to a certain genre and style.

The information system should take into account the stages of analysis of poetic texts. The structure of the system consists of the components listed in the description of the problem statement. The connections in the structure are shown in Fig. 4. The phonetic analysis component is included in the system as a ready-made module, it performs the accentuation (placement of accents) and the transcription of words. The metrorhythmic elements included in the module use the input data obtained at the stage of structural analysis. Taking into account the occurrences of words and collocations in certain poetic texts is associated with the task of compiling the frequency reference books and concordances, and with pragmatic analysis, the genre and style of the work are determined. The components of the user in-

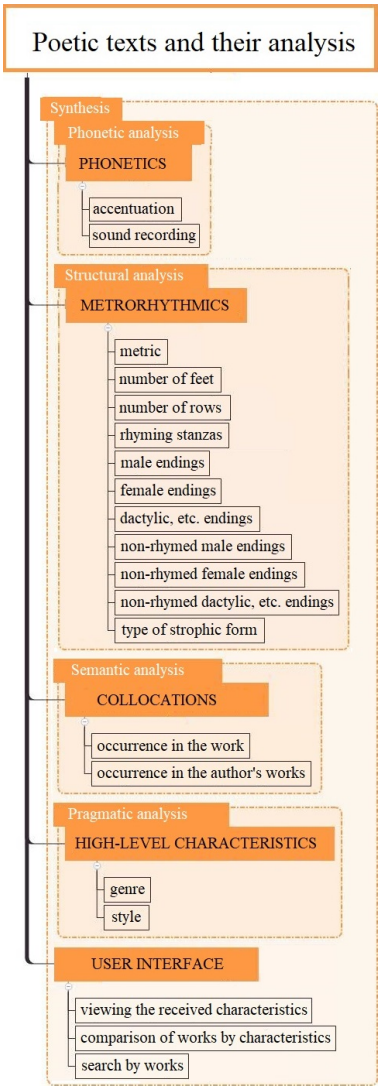

Fig. 4. Structure of the information system

The technical requirements for the information system

Requirement	Priority
Take into account the above-described stages of complex analysis	high
Display metrorhythmic characteristics	high
Display the results of semantic analysis for each work	medium
Develop the functionality of manual data correction by an expert	medium
Implement access to the information system for the end user	high
Provide the logic of the system for repeating values	medium
Design the logic of the system for its further scaling	high
Develop an API (application programming interface) for accessing the information system	high

terface represent the key features for working with the system: viewing characteristics and comparing poetic texts on them, as well as searching for works.

The technical requirements for the information system are shown in Table. For each requirement, the priority of its implementation is set:

- high — the fulfillment of such requirements is necessary;
- medium — the fulfillment of such requirements is desirable;
- low — the fulfillment of such requirements is not necessary (there are no such priority in our task).

Fig. 5 shows the use case diagram. The system was designed taking into account the fact that in the process of operation it will be used by various categories of users and interact with other systems. The following actors were selected:

- a user — a philologist conducting research on poetic texts;
- a programmer-researcher — a person responsible for the most complete implementation of the functionality and operation of the system, as well as for the mathematical processing of the results;
- a system administrator — a person responsible for using special technical means, adding new content, updating data, etc.;
- an external system — one or another subsystem of the project for automating the analysis of poetic texts which exports the results of the researches to the system for a comprehensive display of the results.

The roles of programmer-researcher and user can be combined depending on the specifics of the philological task. In addition, a user, as an expert, can evaluate the data output by the system and, in case of disputes, edit the data in the information system. All actions in the system must be logged by administrator — for example, when importing data from project subsystems, unforeseen technical problems may arise that need to be fixed manually.

Next, we present some scenario of interaction of possible participants in the system (according to the use case diagram diagram). A user of the information system has the opportunity to view information about the works, after conducting an expert assessment and in case of finding disputed data, a philologist user sends a request for access to editing or informs the database administrator about the results of the expert assessment and the decision to change the data. A database administrator checks the exported data from external systems that analyze the poetic texts for technical errors (in case the external system received errors as a result of the analysis). A programmer-researcher has the opportunity to use the API to organize batch unloading of data on the analyzed poetic texts.

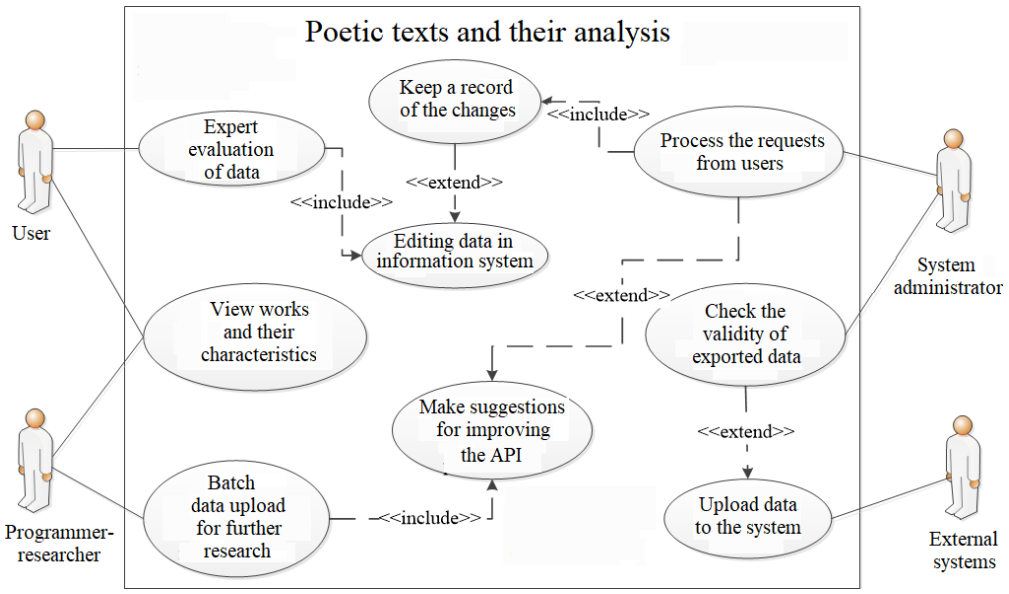

Fig. 5. The use case diagram of information system

The external systems (subsystems), using pre-established interaction protocols, load the results of the analysis of lyrical works obtained during the operation of individual algorithms (including from the subsystem described in [32]). In the future, a variant of interaction is being developed in which subsystems use the data of the described information system as input parameters.

4. Conceptual model of the system

The main part of the information system is implemented in the form of a database that stores both the poems themselves and their characteristics. The certificate of registration of the database was obtained for the database "Russian poetic texts and their complex poetic metadata" [35]. The conceptual ER-model of the system is shown in Fig. 6. To standardize

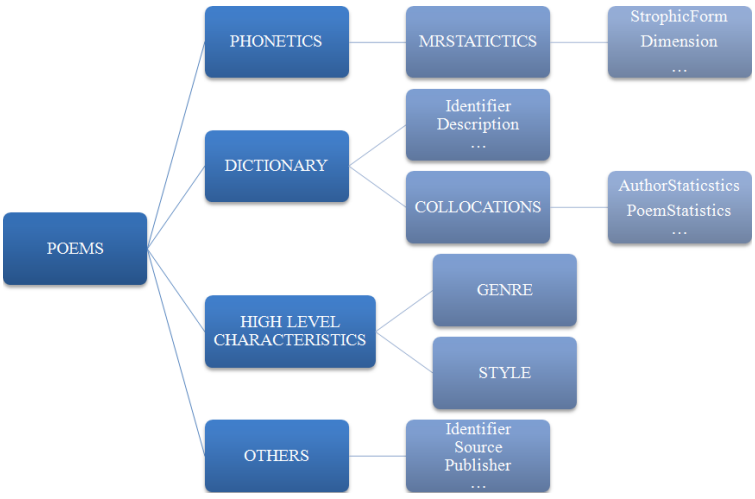

Fig. 6. Conceptual ER model of an information system

the description of metadata, The State Standard R 7.0.10-2010 [36] was used, based on a set of elements of the "Dublin Core" [37] — the standard for the system-wide description of information resources.

The entity POEMS includes.

1. The phonetic characteristics (the entity PHONETICS), on the basis of which the metrorhythmic characteristics (the entity MRSTATISTICS) are determined; the latter, in turn, includes the metadata.
2. The dictionary of words (the entity DICTIONARY) is associated with the entities of "COLLOCATIONS", which are extracted from the text; the key metadata of the COLLOCATIONS entity is the statistics of mentions in the work (PoemStatistics) and in the works of an author (AuthorStatistics) in general.
3. The high-level characteristics (the entity HIGH LEVEL CHARACTERISTICS) include data about genre (GENRE) and style (STYLE).
4. The other characteristics (the entity of OTHERS) are auxiliary and serve to store heterogeneous external information about the works, for example, source (Source), information about the publication (Publisher) and others.

5. Information system interface

To organize the access to the information system, the user interface was designed and developed, accessible through the user's web browser. Database queries are organized in SQL format, what increases the versatility of using the system. The interface is available at `http://db4.sbras.ru/~poems_user/IS`. During the test implementation of the system as the experimental corpus of texts, it was decided to use the cycle of lyrical works by A.S. Pushkin, written in 1830, so-called "Boldino's Autumn", an example is shown in Fig. 7 (an example of how the system works is presented in Russian, since the software system is designed to work with texts in Russian).

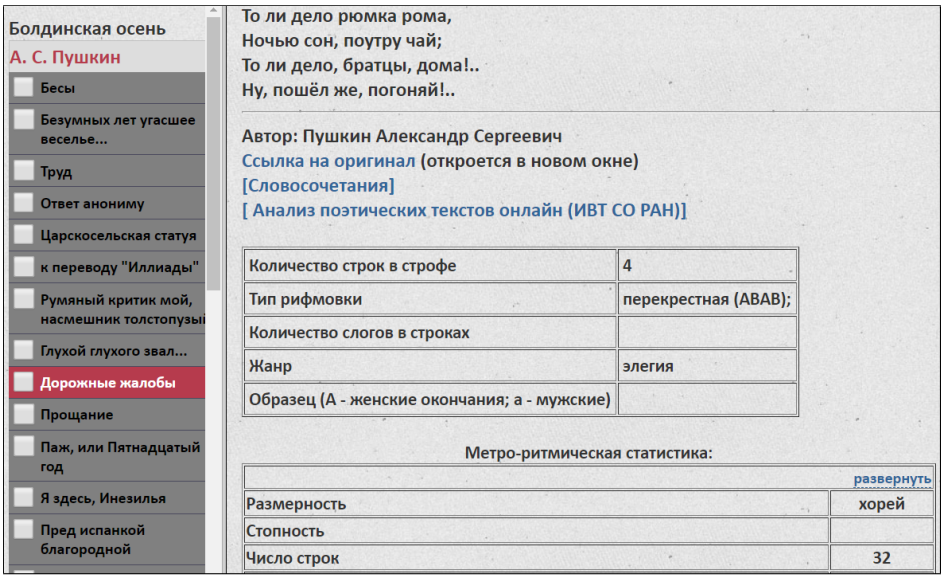

Fig. 7. The fragment of the information system interface displaying an example from the test sample

By selecting the available works of the author (menu on the left), a user gets the opportunity to both view the work itself and get acquainted with its characteristics. For some works, a link to the handwritten original is provided — the resources of the Institute of Russian Literature [38] (IRLI RAS) were used. An additional window for collocations opens when clicking on the link of the same name in the work viewing mode: a user has the opportunity to view the selected phrases and combinations of words, and if there are appropriate access rights to change them (after an expert assessment). The functionality of comparing several works is implemented — a user selects the works necessary for analysis: a summary table of works with key characteristics is displayed for viewing.

6. Methodology

To design the information system for presenting the results of complex analysis of poetic texts in Russian, the RUP design methodology was used [39], which is also used for the design of large-scale systems, using an iterative development model that allows to quickly respond to changing requirements, as well as effectively control the quality of the model being created. The generated requirements for the system and its structural units were checked for compliance with the IEEEStandard 830-1998 [40], which is the part of methodology for compiling specifications of software requirements. The design and formalization of business processes were carried out using the modeling methodologies IDEF0 [34], intended, among other things, for the formalization and description of processes, and having a distinctive feature of the emphasis on the subordination of objects, and BPMN [41], which allows to define complex semantic constructions, creating a standard set of symbols that are understandable to all users and is a kind of link between the design phase of the project and the phase of its implementation. The modeling of the system structure and user roles was carried out using UML notation [42], which uses graphical notation to create the abstract model of the system.

Conclusions: the conceptual description of the software system

The system describes the following types of entities:

1) authors;
2) texts;
3) lexemes;
4) rhymes;
5) metrorhythmic characteristics (factures);
6) genres;
7) styles.

The central component is the system for storing texts and their metadata. The basic element of the description is the word in each of its occurrences in the poetic text. The four-level coding is proposed:

1) ID author;
2) ID poem;
3) ID line in a poem;
4) ID word in a line.

Thus, each word is defined by four indexes.

The system allows to automatically generate the following dictionaries and reference books, the main ones for a philologist.

1) lexical dictionary;
2) language dictionary;
3) concordance;
4) rhyme dictionary;
5) metrorhythmic reference book.

The lexical dictionary stores the words in the initial form. The referring to the initial forms of words allows to get a dictionary of the language. The referring to word forms in the context of words allows to get a concordance.

The information about each rhyme is presented in the form of an n-ary relation, where $n \geq 2$. By referring to this component, an automatic compilation of a rhyme dictionary is organized.

The usage of metrorhythmic and strophic metadata of the corpus of poems makes it possible to organize a metrorhythmic reference book.

Thus, as a result, the conceptual design and implementation of an information system for presenting the results of automated complex analysis of poetic texts, stated in our early work [43], was carried out.

References

[1] **Gitt W.** Ordnung und information in technik und natur. Gitt W. (Hrsg.): Am Anfang war die Information. Gräfeling: Resch KG; 1982: 171–211. (In German)

[2] **Shannon C.E.** A mathematical theory of communication. Bell System Technical Journal. 1948; 27(3):379–423.

[3] Ozhegov S.I. Tolkovyi slovar' russkogo yazyka [Explanatory dictionary of the Russian language]. Moscow: Mir i Obrazovanie, Oniks; 2011: 736. (In Russ.)

[4] **Lotman Yu.M.** Struktura khudozhestvennogo teksta [The structure of a literary text]. Moscow: Iskusstvo; 1970: 384. (In Russ.)

[5] **Magomedova D.M.** Filologicheskiy analiz liricheskogo stikhotvoreniya [Philological analysis of lyric poems]. Moscow: Akademiya; 2004: 187. (In Russ.)

[6] **Taranovskiy K.F.** O vzaimootnoshenii stikhotvornogo ritma i tematiki [About the relationship between poetic rhythm and thematics]. In: Taranovsky K. About poetry and poetics. Moscow: Yazyki Russkoy Kul'tury; 2000: 372–403. (In Russ.)

[7] **Hayward M.** Analysis of a corpus of poetry by a connectionist model of poetic meter. Poetics. 1996; 24(1):1–11. Available at: `http://www.english.iup.edu/mhayward/Metrics/Cormetrics.htm`.

[8] **Lapshina N.V., Romanovich I.K., Yarkho B.I.** Metricheskiy spravochnik k stikhotvoreniyam A.S. Pushkina [Metrical Guide to the poems by A.S. Pushkin]. Moscow, Leningrad: Academiya; 1934: 144. (In Russ.)

[9] **Samoylov D.** Kniga o russkoy rifme [Book about Russian rhyme]. Moscow: Khudozhestvennaya Literatura; 1982: 351. (In Russ.)

[10] Stemmer of the company "Yandex". Available at: `https://tech.yandex.ru/mystem`.

[11] **Shokin Yu.I., Fedotov A.M., Barakhnin V.B.** Problemy poiska informatsii [Problems of information retrieval]. Novosibirsk: Nauka; 2010: 196. (In Russ.)

[12] **Kozmin A.V.** Automatic analysis of verse into the Starling system. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2006". Moscow; 2006: 265–268. (In Russ.)

[13] The project "The Tower of Babel". Available at: `http://starling.rinet.ru/indexru.htm` (In Russ.)

[14] Analiz poeticheskikh tekstov onlayn [The analysis of poetic texts online]. Available at: `http://poem.ict.nsc.ru`. (In Russ.)

[15] **Barakhnin V.B., Kozhemyakina O.Yu., Zabaykin A.V., Hayatova V.D.** Automation of complex analysis of Russian poetic text: models and algorithms. Vestnik NGU. Seriya: Informatsionnye Tekhnologii. 2015; 13(3):5–18. (In Russ.)

[16] **Boikov V.N., Karyaeva M.S., Sokolov V.A., Pilschikov A.I.** About automatic specification of the verse in the information-analytical system. CEUR Workshop Proceedings. 2015; (1563):144–151. Available at: `http://ceur-ws.org/Vol-1536/paper22.pdf`. (In Russ.)

[17] **Pilshchikov I.A., Starostin A.S.** Osnovnye problemy avtomatizatsii bazovykh protsedur ritmiko-sintaksicheskogo analiza sillabo-tonicheskikh tekstov. Natsional'nyy korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy [The main problems of automation of basic procedures of rhythmic-syntactic analysis of syllabotonic texts. National Corpus of the Russian language: 2006–2008. New results and prospects]. Sankt-Peterburg: Nestor-Istoriya; 2009: 298–315. (In Russ.)

[18] **Pilshchikov I.A., Starostin A.S.** The problem of automatic meter recognition: syllabotonics, dolnik, taktovik. Russian poetry: 100-year results and prospects of development. International Scientific Conference Proceedings. November 25–27, 2010. St. Petersburg; 2010: 397–406. (In Russ.)

[19] **Pilshchikov I., Starostin A.** Reconnaissance automatique des m?tres des vers russes: une approche statistique sur corpus. Langages. 2015; 3(199):89–106. (In French)

[20] Zapis' doklada A.S. Starostina "Instrumental'naya sreda dlya raboty s russkoyazychnymi stikhotvornymi korpusami i ikh spetsializirovannoy razmetkoy" [Report by A.S. Starostin "Instrumental environment for working with Russian-language poetic corpuses and their specialized markup"]. Available at: `https://youtu.be/TUWIzxtveNY`. (In Russ.)

[21] **Barakhnin V.B., Kozhemyakina O.Yu., Pastushkov I.S.** Comparative analysis of methods of automated classification of poetic texts based on lexical signs. CEUR Workshop Proceedings. 2017; (2022):252–257. Available at: `http://ceur-ws.org/Vol-2022/paper41.pdf` (accessed 25.09.2018). (In Russ.)

[22] **Bulygin M.V., Sharoff S.A.** Using machine translation for automatic genre classification in Arabic. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2006". 2006: 153–162.

[23] **Loukachevitch N.V., Rusnachenko N.** Extracting sentiment attitudes from analytical texts. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". 2018: 459–468. Available at: `https://www.dialog-21.ru/media/4317/loukachevitchnv_rusnachenkon.pdf`.

[24] **Polyakov A.Y., Pilschikov I.A., Bergelson M.B.** Konkordans k tekstam Lomonosova — kontseptsiya i realizatsiya [Lomonosov concordance — concept and implementation]. Available at: `http://www.dialog-21.ru/digests/dialog2009/materials/html/61.htm`. (In Russ.)

[25] **Polyakov A.Y., Pilschikov I.A., Bergelson M.B.** Konkordans k tekstam Lomonosova [Lomonosov concordance]. Available at: `http://feb-web.ru/feb/lomoconc/abc`. (In Russ.)

[26] **Delmonte R.** Computing poetry style. Proceedings of 1st International Workshop ESSEM 2013. CEUR Workshop Proceedings. 2013; (1096):148–155.

[27] The project SPARSAR. Available at: www.sparsar.wordpress.com.

[28] **Bacalu C., Delmonte R.** Prosodic modeling for speech recognition. Atti del Workshop AI*IA — "Elab.Ling.e Ric." IRST Trento; 1999: 45–55.

[29] The project Metricalizer. Available at: https://metricalizer.de.

[30] **Bobenhausen K., Hammerich K.** Literary metrics, linguistic metrics, and the algorithmic analysis of German poetry using Metricalizer. Languages. 2015; 199(3):67–87. DOI:10.3917/lang.199.0067.

[31] **Barakhnin V.B., Kozhemyakina O.Yu., Borzilova Yu.S.** The development of the information system of the representation of the complex analysis results for the poetic texts. Vestnik NGU. Seriya: Informatsionnye Tekhnologii. 2019; 17(1):5–17. DOI:10.25205/1818-7900-2019-17-1-5-17. (In Russ.)

[32] **Barakhnin V.B., Kozhemyakina O.Yu., Mukhamediev R.I., Borzilova Yu.S., Yakunin K.O.** The design of the structure of the software system for processing text document corpus. Business Informatics. 2019; 13(4):60–72. DOI:10.17323/1998-0663.2019.4.60.72.

[33] **Barakhnin V.B., Kozhemyakina O.Yu., Rychkova E.V., Gladkikh A.S., Pastushkov I.S.** Software for learning to solve problems of classification using of machine learning. European Proceedings of Social & Behavioural Sciences. 2018; (XLIX):106–112. DOI:10.15405/epsbs.2018.11.02.12.

[34] Metodologiya funktsional'nogo modelirovaniya IDEF0. Rukovodyashchiy dokument. [The methodology of functional modeling IDEF0. Th guidance document]. Available at: https://nsu.ru/smk/files/idef.pdf. (In Russ.)

[35] **Barakhnin V.B., Kozhemyakina O.Yu., Borzilova Yu.S.** Russkie poeticheskie teksty i ikh kompleksnye stikhovedcheskie metadannye. Svidetel'stvo o gosudarstvennoy registratsii bazy dannykh No. 2020621889 ot 15.10.2020 g. [Russian poetic texts and their complex poetic metadata. Certificate of state registration of the database No. 2020621889 dated 15.10.2020]. Available at: https://www.fips.ru/ofpstorage/Doc/PrEVM/RUNWDB/000/002/020/621/889/2020621889-00001/document.pdf. (In Russ.)

[36] GOST R 7.0.10-2010 (ISO 15836:2003). Nabor elementov metadannykh "Dublinskoe yadro" [GOST standard R 7.0.10-2010 (ISO 15836:2003). Set of metadata elements "Dublin Core"]. 2011-07-01. Moscow: Standartinform; 2011: 12. (In Russ.)

[37] The Dublin Core metadata initiative. Available at: http://dublincore.org.

[38] Elektronnye publikatsii Instituta russkoy literatury (Pushkinskogo Doma) RAN [Electronic publications of the Institute of Russian Literature (Pushkin's House)]. Available at: http://lib.pushkinskijdom.ru. (In Russ.)

[39] **Maglinets Yu.A.** Analiz trebovaniy k informatsionnym sistemam [The analysis of information system requirements]. Available at: https://ivan-shamaev.ru/wp-content/uploads/2013/06/Information-systems-analysis-and-requirements-analysis.pdf. (In Russ.)

[40] IEEE recommended practice for software requirements specifications. Available at: https://ieeexplore.ieee.org/document/720574.

[41] Business process model and notation. Available at: http://www.bpmn.org.

[42] Unified modelling language. Available at: http://www.uml.org.

[43] **Barakhnin V.B., Kozhemyakina O.Yu.** About the automation of the complex analysis of Russian poetical text. CEUR Workshop Proceedings. 2012; (934):167–171. Available at: http://ceur-ws.org/Vol-934/paper27.pdf. (In Russ.)

Вычислительные технологии, 2022, том 27, № 2, с. 122–137. © ФИЦ ИВТ, 2022 ISSN 1560-7534
Computational Technologies, 2022, vol. 27, no. 2, pp. 122–137. © FRC ICT, 2022 eISSN 2313-691X

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

DOI:10.25743/ICT.2022.27.2.010

Концептуальное проектирование программной системы автоматизированного комплексного анализа поэтических текстов

О. Ю. Кожемякина

Федеральный исследовательский центр информационных и вычислительных технологий, 630090, Новосибирск, Россия

Контактный автор: Кожемякина Ольга Юрьевна, e-mail: olgakozhemyakina@mail.ru

Поступила 31 марта 2022 г., принята в печать 7 апреля 2022 г.

Аннотация

Процесс исследования поэтического текста состоит из анализа уровней структуры стиха, модель которой находится в строгом соответствии с моделью информационного сообщения. Каждый уровень рассматривается как самостоятельная смысловая единица с последующим связыванием полученных данных с другими элементами структуры. Каждый модуль программной системы автоматизированного анализа поэтических текстов относится к одному из уровней анализа текста: структурному, семантическому, прагматическому. Структурный анализ поэтического текста связан с выделением его метроритмических характеристик. В рамках семантического анализа находятся исследования по выделению смысловых конструкций из произведения; прагматический уровень включает в себя исследования по автоматическому определению высокоуровневых характеристик поэтического текста, таких как жанр и стиль.

В работе представлен процесс концептуального проектирования и реализации программной системы автоматизированного комплексного анализа поэтических текстов. Термин “информационная система” используется как обозначение соответствующего компонента программной системы автоматизированного комплексного анализа поэтических текстов.

На этапе проектирования сформулированы задачи, которые призвана решать информационная система представления результатов комплексного анализа поэтических текстов, а также изложены требования в порядке приоритета для программной системы в целом. Разработанная система объединяет в себе разнородную информацию о результатах анализа поэтических текстов, полученных на каждом из уровней представления. Исходя из потребностей потенциальных пользователей решены следующие задачи: выполнено описание внешних взаимодействующих элементов системы; сформулировано концептуальное описание системы; разработан интерфейс для доступа к хранилищу информационной системы; реализован графический интерфейс информационной системы. Результаты работы показали принципиальную возможность интеграции компонентов системы. Создан удобный для филологов инструментарий автоматизации комплексного анализа поэтических текстов на русском языке.

Ключевые слова: обработка естественного языка, проектирование программной системы, информационная система анализа текста.

Цитирование: Kozhemyakina O.Yu. Conceptual design of the software system for automated complex analysis of poetic texts. Computational Technologies. 2022; 27(2):122–137. DOI:10.25743/ICT.2022.27.2.010.