Advances in Social Science, Education and Humanities Research (ASSEHR), volume 312
International Conference "Topical Problems of Philology and Didactics:
Interdisciplinary Approach in Humanities and Social Sciences" (TPHD 2018)

Technology of the Formation of Frequency Dictionaries for Quantitative Analysis of Tajik Literature

Umarov Makhmud
Department of Informatics and Information Systems
Russian-Tajik Slavonic University
Dushanbe, Republic of Tajikistan
m_umarov@mail.ru

Takhmuradova Diana
Department of Theoretical and Applied Linguistics
Russian-Tajik Slavonic University
Dushanbe, Republic of Tajikistan
diana.rashidovna.00@mail.ru

***Abstract*—The article provides a brief description of the developed tools for the formation of frequency dictionaries. The main stages of the technology of the formation of frequency dictionaries and performing statistical analysis of poetic material are determined. To create a human-computer system that implements this tool, the authors propose the approach based on the concept of the algorithm of dialog operations. The need to develop a software package for the formation of frequency dictionaries arises from the fact that the vast heritage of the Tajik-Persian classical literature is still not properly investigated from the standpoint of statistical analysis. The use of statistical methods gives the best results in stylistics, which can dispose of subjective assessments with the help of counting and strict systematization of the material. During the study of Tajik-Persian classical poetry, it is of great interest to establish patterns in other parameters of poetic material as well apart from attribution, which is undoubtedly one of the most important tasks of stylometry. A significant part of the research process in this direction is the routine work of processing textual material. In this regard, there is a natural need to create such a toolkit that automates all the basic routine work of the research process.**

Keywords—frequency dictionary; technology; quantitative method; algorithm; program; automation; dialog operations

I. INTRODUCTION

1. Problem statement

The study of poetic and non-poetic styles can be based on statistical methods. Moreover, it is possible to conduct relevant research based on specific forms (parameters) that language laws take in texts of various styles. In such cases, quantitative linguistics conducts research in stylistics: one of the final goals is to prove the existence of a stylistic phenomenon as objectively as possible, in at least one area of action, referring to the operation of language law. If poetic texts are studied, then methods of quantitative linguistics form a subdiscipline, which is called "Quantitative study of literature" or stylometry. Modern stylometry is usually built on the use of computers to the analysis of texts [1-4, 5; pp. 760-774].

The need to develop a software package for the formation of frequency dictionaries arises due to the fact that the vast heritage of the Tajik-Persian classical literature has not been properly studied from a position of statistical analysis. The existing separate works [6] do not cover the entire volume of this heritage. The use of statistical methods gives the best results in stylistics, which can dispose of subjective assessments with the help of counting and strict systematization of the material [7].

Another important task for the development of such technologies is the creation of the National Tajik Corpus. In our country corpus linguistics undergoes the period of formation; therefore the creation of the National Tajik Corpus is the most important of the tasks. Under these conditions, the creation of corpuses of texts is a real breakthrough in this area and it is quite obvious that the corpus software itself and dictionaries based on the corpus may have some errors. It is obvious that in order to solve such problems as text markup, the great amount of time and efforts of specialists creating the necessary programs taking into account the peculiarities of the Tajik language.

In addition, a significant part of linguistic tasks requires the identification of historical changes in the functioning of linguistic phenomena — for example, the changes in the meaning of words, the frequency of use of certain syntactic structures, etc. Moreover, the corpuses of such famous Persian-Tajik poets as Rudaki, Firdavsi, Mavlono and others belong to the corpus of poetic texts, since they include poetic works. Accordingly, in addition to the usual semantic and morphological markup, special poetry markup should be provided. Thus, it will be possible to search for texts written by various poetic forms, according to given parameters.

For finding and eliminating such shortcomings and errors, feedback communication is undoubtedly important. The feedbacks of users of these cases and dictionaries will improve the work in this direction.

A group of scientists of Masaryk University (Czech Republic), together with the Tajik scientist Gulshan Dovudov, relatively recently built a corpus of the modern Tajik language, containing more than 50 million words. The authors of the project noted that all the texts of this corpus were taken from the Internet.

The authors of the project also note that two partial corpuses were combined, and the result was checked for duplication using the Onion program. As a result, the Tajik language corpus was obtained, consisting of more than 50 million Tajik words and more than 60 million tokens (words, numbers, punctuation marks) [8].

The research and applied work in this area are continuing, perhaps there are other significant results that, in our opinion, should be coordinated, since projects of this kind should have governmental status.

2. Methods. Technologies

Modern programming technologies allow developing various applications that find their implementation in various subject areas. The methodological basis of these technologies is presented by [9;10]: the theory of algorithms and algorithmic methods; structured programming methods; object-oriented programming techniques; visual modeling techniques and information exchange and data access technologies.

Typically, when designing an application, problems may arise not only in the development of an effective software package, but also in providing access to heterogeneous data sources. Fortunately, nowadays there exist such technologies as DAO (Data Access Objects), RDO (Remote Data Objects) and ADO (ActiveX Data Objects) that give access to very diverse data stores.

In our opinion in the course of the development of office applications, the DAO technology with the Jet database engine is the most effective one. Since in a huge variety of programming languages Visual Basic for Application (VBA) is unique and natural from the point of view of developing office applications, it is therefore considered more appropriate and accurate to use a combination of Jet and DAO.

II. Results and Discussion

1. Program complex description

The main stages of the process of the formation of a frequency dictionary and performing statistical processing of poetic material are as follows:

1. The preparation of material.
2. The process of material reading and the formation of a table in database
3. The pretreatment of material and normalization of lexical elements.
4. The formation of concordance and the provision of the possibility of editing.
5. The editing of concordance in order to determine additional parameters.
6. The statistical analysis in accordance with the specified parameters.

It is obvious that this software package is a human-computer system which provides interaction between various software environments (text editor, DBMS, spreadsheet table).

The description of the models of the elements of the software package (PC) is made on the basis of the UML (Unified Modeling Language) standard approved by the OMG (Object Management Group) in 2004 [11]. To achieve this, the IBM Rational Rose, Enterprise Edition 7.0 tools were used to track the development, maintenance and operation of a software package throughout its life cycle. The possibilities of practical use of IBM Rational Rose are presented in the works [12; 13].

In order to ensure the completeness of descriptions of models, a set of requirements for the final product has been formed, primarily based on the functions that a frequency dictionary should perform and the tasks that arise in the process of its formation. Therefore, the main functional requirements for the developed project of a frequency vocabulary software complex are formulated as follows:

- The developed project of a frequency vocabulary software complex should provide processing and storage of the results of poetic works of various authors and various genres;
- The processing of the source material is carried out in several stages:

1. Primary processing - the normalization of lexical elements;

2. The formation of a list of lexical elements with an indication of their location in the material with line accuracy;

3. The determination of the length (number of letters) of lexical elements;

4. The determination of the frequency of lexical elements;

5. The sorting of lexical elements in accordance with the specified attribute (length, alphabet, occurrence, frequency);

• The processing results at any stage should be stored separately from source material in order to restore it in case of software or hardware failure;

• The developed project of a frequency vocabulary software complex should provide the ability to edit in the following modes: preliminary, intermediate, final;

• The developed project of a frequency vocabulary software complex should ensure the preparation and printing of the following output documents:

1) The list of lexical elements in accordance with the selected attribute, fully or partially, with an indication of the location of a word in the material or without it;

2) The list of the changes made as a result of preprocessing and their initial form;

3) The results of statistical processing of material data in the form of tables, charts, graphs and comments.

In the framework of this research we are limited to the description of the material preprocessing algorithm, the block diagram of which is shown in Figure 1.

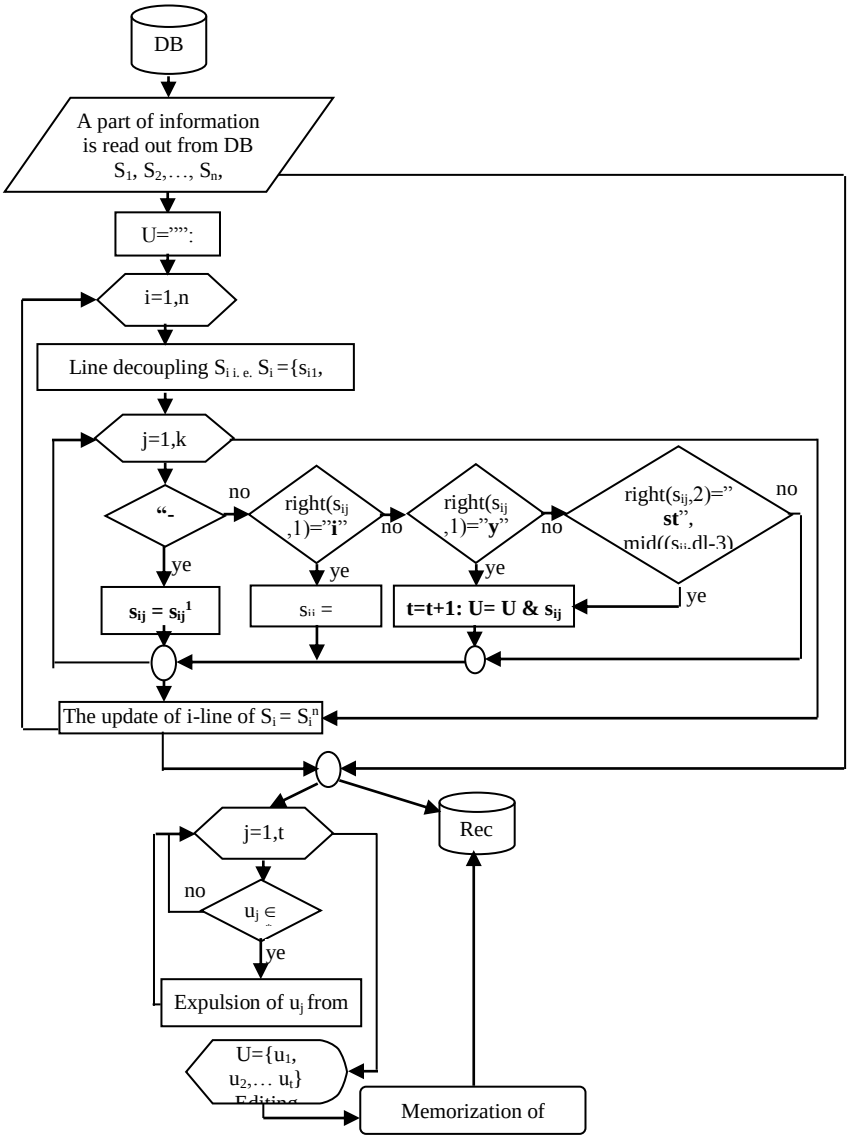

Fig. 1. Block diagram of the preprocessing algorithm.

The UML notation system allows describing the dynamic behavior of a software package and its static structure. Modern CASE technologies (in particular, Rational Rose) contain various means of generating code, and the possibility of choosing a programming language is also provided. However, we did not use this feature of the Rational Rose program according to the following considerations:

1. First of all it is due to the inability to automate the entire process of the formation of a frequency dictionary. Moreover, it is not recommended because in addition to the formal routine work of this procedure, there are a large number of informal problems, the solution of which completely depends on expert knowledge. These are the tasks of semantic and etymological interpretation of words, separation of homonyms, determination of belonging to one or another part of speech, etc. Based on this, there is a natural need to provide end-users with a wide range of possibilities for manipulating source, intermediate and resulting data.

2. The application must be based on an accessible platform for end users. Nowadays such a platform is the Office software, since almost all of its individual components are used by almost all those who perform a task on a personal computer, and by some estimates 90% of organizations use the possibilities of Office software by only 10% [14]. Therefore, the mechanism of data manipulation must be provided with the help of these programs through the wide use of their functions. In this regard the problem of integrating various applications of MS Office is relevant.

The interface of the program is written in Visual Basic. Previously, a special program block reads out material prepared in a text editor and forms a database in MS Access. **Next,** interacting with MS Access programmatically, the necessary adjustments are made in the database. In the same

block, a list of lexical elements in the studied material is displayed through a dialog box, some of which are marked for memorization and learning. Then the first version of the list is formed, which is displayed on the MS Excel. Excel is chosen because of the convenience of editing lists and the ability to perform computational procedures. Interacting through the system interface, the user receives various options for lists. After the formation of the list of the frequency of lexical elements, which is simultaneously drawn up in the form of a concordance (index of usage) in a separate file for further use in the environment of text editors, the possibility of editing is provided. Depending on the task, the parameters of lexical elements are determined. A separate group of macros is intended for statistical study of the material according to certain parameters.

In this regard, the advantage of this approach, despite its labor intensity, over the automation of code generation based on CASE technologies, becomes obvious. It does not mean that we completely refute the benefits of using CASE - technology and all the useful consequences resulting from it. However, when solving this problem, due to the fact that the studied area contains a great many unclear and unusual methods of the formation of new elements, the validity of this statement becomes obvious. The theoretical basis of the approach used in this work was developed in the work [15] in order to create problem-oriented CAD subsystems and its modified version was implemented in the work [16].

The basis of this approach is formed by the concept of the algorithm of dialog operations (ADO), which is understood as a set of logically related operations performed by a man and computer, denoted as graph-schemes for processing, entering and outputting data when solving specific applied problems implemented as problem-oriented software complex. All sets of ADO operations are divided into two main classes: formal operations on data performed by a computer automatically, without participation of a person in this process; informal operations of converting the data inserted into a computer by a person (Fig. 2).

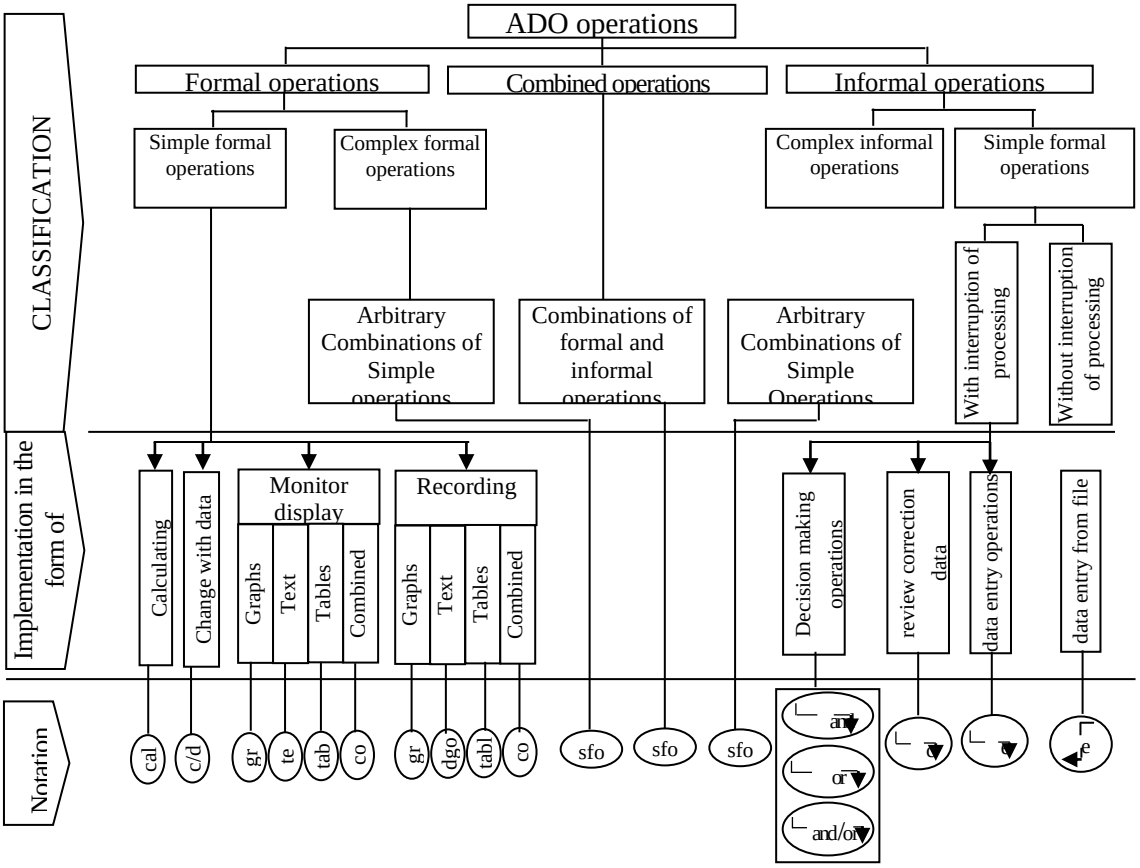

Fig. 2. Block diagram of the preprocessing algorithm.

Simple informal operations are divided into two types: with the interruption of the processing and without interruption of it. Simple informal operations with the interruption of the process of problem solution require the presence of two initial steps: 1) issuing an introductory informational message; 2) issuing, in any form, invitations to work and putting a task in a sleep mode.

The difference between a simple informal operation and previous operation is that it lacks moments of putting a task in a sleep mode and issuing a command to continue solving a problem. In this case, a man and a computer work simultaneously in the modes: man-preparation of information, computer-display of information on a monitor and processing of the information prepared by a user. In the UML notation system, the decision and flow synchronization charts are used in order to achieve it.

According to the definition of ADO and for its creation in relation to the problem-oriented software package it is necessary to have the following information:

1. The formulation of the purpose of the software package and the definition of the main requirements for it;

2. The models and methods for the solution of the studied problems;

3. The algorithms for the processing and block diagram of it;

4. The models of the presentation of various aspects of the software package and their description based on a standardized notation system;

5. The lists of input and output data flow, their forms and limits;

6. The list of output documentation with a specific form of presentation of the results;

7. The presentation of dialogue procedures at three stages: setting the source data, setting up the task, choosing a solution method, etc.; reviewing current information, correcting the source data and making the required changes, saving intermediate results, etc. and reviewing the results, designing organization of issuance of documentation, etc.

The significance of the first two indicated stages is especially great during the course of the development of this type of software, the end users of which are domain specialists. Therefore, the use of technology based on ADO allows developing a software package that can be convenient, accessible and understandable to the end user, regardless of the application.

III. CONCLUSION

In the systems of automatic information processing, semantic-syntactic analysis of texts is carried out with the purpose of a formalized presentation of their structure — the identification of semantic units and the establishment of links between them. The structure of the texts can be interpreted in different ways and described in various formalized languages. The specific purposes and results of the analysis can also be different. A sentence is traditionally considered as the main structural unit of a text. The sentences appear in the text not in isolation from each other, but in close semantic connection. This connection is based on the mental images of those concrete or abstract objects (situations, phenomena) that a person has in his mind when he generates a text. The images of these objects have a certain structure. In addition, they are additionally structured by a person when describing them in natural language. Accordingly, a text is structured.

The syntactic structure of texts is usually described in terms of classes of words and their relations. At the same time, parts of speech (noun, adjective, verb, adverb, etc.) accompanied by grammatical information characterizing specific forms of words (for example, gender, number, case, person, etc.) can act as classes of words. The relation of direct domination with varying degrees of differentiation can act as the relations of classes of words.

The analysis of the syntactic structure of a sentence must be performed on the basis of the information about the words obtained at the stage of morphological analysis. At the same time, each word form of a text is assigned a corresponding symbol of a grammatical class and a set of grammatical features. These are parts of speech. For example the formation of parts of speech is the following: “oftob” is a noun, and “oftobi” is an adjective, or the verb “huftan” is based on a noun “hob”, etc.

Therefore, at the first stage of the analysis, the need for statistical processing of qualitative features of a text becomes obvious. Qualitative signs of grouping option are such signs that do not contain either a quantitative assessment of an option, or the possibility of their ranking. An example can be presented by the grouping of word forms into semantic or grammatical classes, or the location of phonemes, based on the hierarchy of differential features. In these cases, the grouping option, selected by quality attribute, is in their classification by the gradation of this feature.

One of the first frequency dictionaries created using this technology was the Hafiz frequency dictionary. The study of the material was carried out on the basis of determining the statistical laws of the following qualitative characteristics:

1. The grouping of word forms by grammatical features. In other words, word forms are grouped according to their parts of speech: noun, verb, pronoun, adjective, adverb, numeral, prepositions, appeal and exclamation;

2. The grouping of word forms by language and other signs. Language signs are: Tajik word, Arabic, Greek, and Turkic. Other signs are: religious, geographical, astronomical, and the name of a person.

The research and applied works in this area are continuing; starting in 2019, we are performing a large project in order to create the National Tajik Corpus and the above mentioned developments and technologies will form the basis of work in this area.

Acknowledgements

The authors express their gratitude to the Russian-Tajik (Slavonic) University for financing the research under the University Development Program for 2018.

References

[1] G.Ya. Martynenko, “Fundamentals of stylometry”, Leningrad: Publishing House of Leningrad State University, 1988, 176 p.

[2] F. Can, J.M. Patton, “Change of writing style with time”, Computers and the Humanities, vol. 38, No. 1, pp. 61-82, 2004.

[3] J. Hope, “The Authorship of Shakespeare’s Plays”. Cambridge, Cambridge University Press, 1994.

[4] A.A. Kenny, “Stylometric Study of the New Testament”, 1986.

[5] V.V. Levitsky, “Quantitative methods in linguistics”. New book, Vinnitsa 2007. Reinhard Köhler: Synergetic linguistics. In: Reinhard Köhler, Gabriel Altmann, Rajmund G. Piotrowski (Hrsg.): Quantitative Linguistic – Quantitative Linguistics. Ein internationales Handbuch. de Gruyter, Berlin/ New York 2005, pp. 760-774.

[6] K.-H. Best, O. Rottmann, “Quantitative Linguistics, an Invitation”. RAM-Verlag, Lüdenscheid, 2017.

[7] R.G. Piotrovsky et al., "Mathematical linguistics: Proc. manual for ped. Institutions", Moscow: Higher school, 1977, pp.64-67.

[8] M.-N. O. Osmanov, "Unsuri Frequency Dictionary". Moscow: Science, 1970, p. 3.

[9] G. Dovudov, V. Suchomel, P. Smerk, "POS Annotated 50M Corpus of Tajik Language", In Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages, Istanbul, Turkey, 2012, p. 93-9S.

[10] Z.D. Usmanov, G.M. Dovudov, "Creating the collection of Tajik language", Dushanbe: Donish, 2014. 109 p.

[11] M.A. Umarov, M.D. Kasymova, "Modern technologies of access to databases in programming issues", Issues of ICT resources in education, Tr. scientific and practical conf., Dushanbe, 2006, pp. 114-121.

[12] A.M. Vedrov, "Workshop on the design of software economic information systems", Proc. manual. – 2nd ed., Moscow: Finance and Statistics, 2006.

[13] T. Kvatrani, "Rational Rose and UML. Visual modeling". Per. from English, Moscow: DMK Press, 2001.

[14] A. Garnaev et al., "Microsoft Office 2000". Application Development. - SPb: BHV, 2000. p.11.

[15] G.S. Nesterenko, V.I. Chechulin, I.A. Gern, "Technology for creating software for problem-oriented CAD subsystems", Methods and tools for design automation: Proc. works. Issue 2 M., 1986, p.3-17.

[16] M.A. Umarov, M.Ch. Yusupov, "The technology of designing dialogue systems for solving problems of programming crops", Coll. Applied Mathematics, Issue 3, Dushanbe, 1991, pp. 131-138.