EDUCATION AND QUALITATIVE RESEARCH: FREQUENCY ALGORITHMS AND THE ARABIC TEACHING MATERIALS

Oleg REDKIN
Faculty of Asian and African Studies, Laboratory for Analysis and Modelling of Social Processes,
St Petersburg State University,
199034, 11, Universitetskaya nab., St Petersburg, Russia

Olga BERNIKOVA
Faculty of Asian and African Studies, Laboratory for Analysis and Modelling of Social Processes,
St Petersburg State University,
199034, 11, Universitetskaya nab., St Petersburg, Russia

ABSTRACT

Modern universities are both centers of education and research, i.e. their mission is to build up high skilled specialists from the one hand and to discover and disseminate knowledge from the other. While the availability of highly skilled specialists is one of the preconditions of every sophisticated scholar research, in its turn, the effectiveness of the teaching process greatly depends on the use and application of the scholar investigations results and technologies often developed in the next door department or lab at the same University. The above mentioned issue is also true when talking about the humanities in general and foreign languages acquisition in particular. The traditional attitudes to the development of textbooks, vocabularies, reference grammar books, etc., often ignore objective data and the results of scholar researches and, as a rule, are based on subjective assessments of the instructors. Meanwhile, implementation of the results of scholar linguistic researches in the teaching process would add to its effectiveness and raise its standards. Thus the authors of the current research have developed algorithms of frequency ranking of words on the basis of the computer analysis of large volumes of graphic texts (including ca. 1 mln. units) and compiled a list of the most frequently used words in Arabic. As a result it allowed using the most frequent vocabulary to compile textbooks, grammar and reading tests, dictionaries, etc. The principle of the implementation of the most frequent lexical units from this list is widely used in the process of teaching of Arabic at St. Petersburg State University, for example in building bilingual dictionaries, on-line course of the Arabic language, teaching books and tests as well.

Keywords — research, education, Arabic, language, frequency.

1. INTRODUCTION

There is a high level of interaction between science and education that often has a character of informal relations. This kind of interaction is implemented primarily in the areas of training of personnel, joint projects, practical implementation of scholar research, etc. These integration processes cover a wide range of different activities and which are implemented in various forms.

For many years the Department of Arabic Studies, St. Petersburg State University has been conducting researches in the field of computer analysis of hand-written Arabic documents, structural analysis of the Arabic texts, and optical character recognition. Results and outcomes of this kind of researches have been integrated into the educational process. For example, in the course of Arabic the results of research on the experimental phonetics are taken into account, especially those related to the phenomena of stress, articulation of consonant phonemes, intonation, etc.

Knowledge of the language of original texts is one of the most important components for the training of skilled specialists in the field of Arabic and Islamic studies. In most of the academic curricula of the departments of the Middle Eastern studies, Arabic is considered as a medium of understanding, expression, communication, and research that allows operating with the original sources, and documents. Each of these spheres of language usage requires a profound knowledge of necessary vocabulary units and information about the rules of its compatibility (i.e. syntax), and variety of word-forms (morphology). Besides that, one should be aware of the contextual and situationally conditioned use of word forms and word combinations.

One of the components of this kind of process is equipment, materials and teaching environment, which includes textbooks, manuals, reference books, and training programs. Among the problems in the course of teaching a foreign language is the preparation of textbooks which convey the peculiarities of a real language within real linguistic contexts, and which represent the most frequent words, phrases, grammar forms most regularly used in real situations.

It is necessary to define the most relevant linguistic information, in this case - the most frequent lexical units, as well as word combinations used in the communication process. Meanwhile, subjective assessments based on personal experience can in most cases not fully reflect the real picture or may even confuse students. The solution to it would be the allocation of the most common vocabulary and the most relevant word forms, and the formation on their basis necessary teaching materials. One of the ways to compose the most effective textbooks and to raise the effectiveness of the language study is to include into it the most relevant vocabulary goal was to find out the most frequent words.

2. FREQUENCY ALGORITHM AND PECULIARITIES OF ARABIC

Each text may be considered as a sequence of linguistic units set in accordance with the internal logic of text. In addition, the structure of words is determined by strict rules of morphology and their linear order is determined by the laws of syntax. There are common characteristics typical for all kinds of texts, i.e. compatibility of its elements with each other in accordance with its internal logic, i.e. the rules of grammar. At the same time, each text reflects the personality of its author, and includes a set of individual peculiarities, for example, such subjective factors as age, gender, education and regional peculiarities of the language which also influence its characteristics.

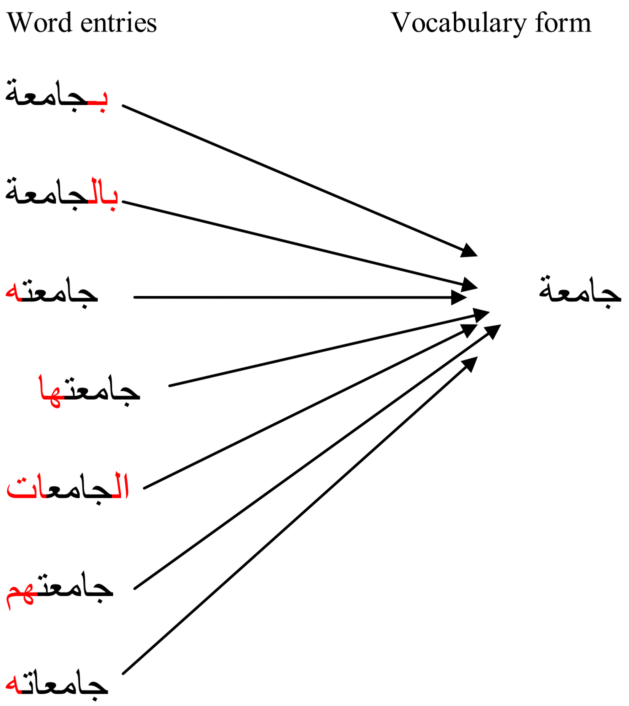
Among the current approaches to the analysis of the material, there are objective methods of mathematic analysis that do not depend on the researcher's evaluations and attitudes, experience and qualifications, but first and foremost rely on the objective data, in this case - qualitative and quantitative.

In the same way, may be carried out analysis of the construction of individual components - i.e. words and phrases, included into the text. One of these indicators is words frequency index.

Arabic belongs to the Semitic language family, in which semantics of most the words is conveyed by three consonantal roots and the formation of grammatical forms takes place mainly due to the change in internal vocal flexion - variations in the quantity and quality of vowels while keeping the invariance of consonant radicals. In addition, the form-building can take place by attaching suffixes, prefixes, infixes, and endings which makes the problem of lemmatization of the Arabic language extremely complex.

When composing a ranked list of words, it was necessary to reduce them to a dictionary form that required the elimination of elements that are written together with the word - suffixes, affixes, prepositions, and various types of endings, i.e. 'to get rid' of ‘irrelevant’ elements.

Fig. 1. Elimination of the graphically connected elements of the word جامعة "university".



Thus following ‘extraction’ of the vocabulary form of the words it was possible to compose the frequency lists of the most relevant lexical elements and to prepare on its basis necessary reading texts, drills,and tests.

3. THE STATISTICAL STUDY OF THE QUR'AN AND ACADEMIC CURRICULA

The language of the Qur'an incorporated many morphological features and vocabulary of the dialects of pre-Islamic Arabia and South Arabian languages. It greatly influenced the formation of the classical Arabic language and facilitated its spread far beyond the borders the Arabian Peninsula along with the spread of the new world religion – Islam, that in its turn largely determined the primacy of the Arab-Muslim culture and science in the early Middle Ages.

For almost 14 centuries the text of Qur'an has been an object of thorough research carried out by Muslim scholarship which ranged from the interpretation of its terminology to the issues related to its recitation and formal characteristics of its text – i.e. a number of entries of individual characters and words. Today the phenomenon of the Holy text is also in the focus of attention of the academic community, who consider it as a unique phenomenon with an indivisible unity of the form and content, which has had a great impact on the history of civilization. The study of the Qur'an is also in the core of educational programs of both traditional Muslim universities of and the secular ones. The basic difference between these two systems of study is in the attitude: in religious universities–most and foremost as the Holy book, while in secular universities it is considered as text with cultural and historical importance.

As a rule, the present studies of the Qur'an are closely related to the curricula of undergraduate and Master's programs. They consider peculiarities of the text, its structure, vocabulary, and phonetics, as well as the impact it has exerted on the religious, historical and cultural process. When studying the features of style, it is important to understand what meanings were initially conveyed by the words, in order to understand their original semantics, on the basis of the reconstruction of initial morphological models, and their role in the syntactic structure of the text. One has to consider and analyze scattered facts and information from various sources. In this regard, traditional methods of research are insufficient and require verification based on objective methods, primarily the instrumental analysis and data-statistical and cluster analysis of the text [1] should be implemented using computer technologies. Since “the learning strategy suggested by the software is to start with the more frequently repeated words since Qur’an contains many repeated words” [2], from the linguistic point of view [3] it necessary to focus the educational process on the most frequent elements included into the text and their interpretation.

In this regard, necessary information related to frequency entries is provided by such an important tool as The Qur'anic Arabic Corpus, developed by scholars at the University of Leeds. The Qur'anic Arabic Corpus, “an annotated linguistic resource which shows the Arabic grammar, syntax, and morphology for each word in the Holy Qur'an” [4] allows to search necessary word-forms by their roots as well as to determine Lemma frequency and provides morphological search for necessary words, and allows to trace ontology of Qur'anic concepts and to acquire information about the Qur'anic grammar. The Corpus may be considered not only as a reference database but also as a scholar research, which has its special value.

Another example of this kind of project is "Morphology of the Qur'an" – a research project which is currently carrying out by scholars at St. Petersburg State University, Russia[1] and which is aimed at the study of the morphological structure of the Qur'anic text, particularly, frequency of forms entries which is important for an adequate understanding of their semantics and contextual usage. The results of this kind of researches are also used when delivering lectures and seminars.

4. MULTIDISCIPLINARITY AND QUALITATIVE RESEARCH

Besides offering tuition in various subjects modern universities provide facilities for scholar research that gives natural opportunities for personal growth as well as intellectual and professional development for both students and faculty.

Among the advantages of modern universities one can mention bringing together experts in different fields of knowledge within the framework of the same educational institution. As a result such kind of 'common umbrella' improves information sharing between research and education departments, provides each of them with new data and scholar ideas which raises the effectiveness of their work.

One of the most eloquent examples of the implementation of the achievements of science in the educational process is the special course "Statistical Data Processing" that was designed for students enrolled in the basic undergraduate program at the Department of African and Asian Studies, St. Petersburg State University (program "The History and Culture of Islam").

The authors and lecturers of the course are scholars from the Department of Mathematics and Mechanics at the University and are also known for their innovative researches and developments in the field of Machine Learning and Data Mining. [5] These ideas are discussed in the course, and unlike other similar programs, it is specially adapted for the targeted audience and the goals of the above mentioned program. Thus the course is focused on the mathematical analysis of the Arabographic texts and presupposes usage of the acquired practical and theoretic skills in the future researches carried out by students themselves.

The teaching course provides students with complex vision of the major stages of qualitative i.e. statistical research, including methods of collecting, processing and presenting of the collected data, as well as considers approaches and solutions of the most typical problems related to qualitative and quantitative evaluations in the context of cultural and historical studies. Students are also provided with such knowledge as measures of communication between data series, correlation of data, their clusterization and Data Mining.

As for the Data Mining, it is generally implemented in the analysis of homogeneous groups of elements in certain data sets. For example, corresponding procedures are applied in statistics, pattern recognition and machine learning. In the case of Islamic Studies these procedures are used in frequency analysis of texts, optical character recognition (OCR) of Arabographic documents, and authorship authorization. [6] Besides that the so called "Bag of Words" method is also often used for the analysis of the text corpora.

[1] Project number 2.15.54.2018

The course also considers the main types of measurements as well as quantitative methods and their assessment in the study of cultural and historical phenomena.

5. CONCLUSIONS

The effectiveness of educational process greatly depends on the introduction into it the results of scholar research. Since the teaching materials are one of the vital elements of the learning environment, along with necessary infrastructure such as electronic equipment, classrooms, skilled personnel, etc. In order to raise the efficiency of the teaching of foreign languages it is necessary to make textbooks and dictionaries more user-oriented, so they should include the most relevant linguistic material.

The texts in any language may be considered as a structured system which may be an object of formal analysis. Despite the complexity of the lemmatization of the Arabic language, modern methods make it possible to identify the most frequent vocabulary, word combinations and morphological models that can be used in the compilation of teaching materials. It is also possible to make text textbooks regionally oriented which presupposes introduction of vocabulary and phrases typical to a particular region or local vernacular.

The mentioned above is true when talking about the study of the language of Qur'an – the most important text in the in the programs of Islamic studies and the standard of the Classical Arabic Language.

The central tenets of this particular type of education besides solid knowledge of the teaching subject should include implementation of the latest methodologies and consider results of scholar researches in that particular as well as other related areas of knowledge which empower students to solve future challenges. They need to be able to put up together discrete ‘chunks’ of knowledge and various facts, and to move from scattered fragments of picture to the general vision of phenomena and to get rid of the constrains imposed by limited vision of the problem. In order to do it is necessary to improve the content and structure of the curricula and to take into consideration results of latest results in related areas of knowledge.

6. ACKNOWLEDGMENTS

The funding no. 15.61.2219.2013 from Saint Petersburg State University (Russia) to support this research project is gratefully acknowledged.

7. REFERENCES

[1] S.K. Safi, "Statistics in Noble Qur'an", **Proceedings of the 2013 International Conference on Information**, **Operations Management and Statistics (ICIOMS2013)**, Kuala Lumpur, Malaysia, September 1-3, 2013.

[2] R.-Y. Raja-Jamilah, M.-Y Zulkifli., Z. Roziati, B. Mohd-Sapiyan, "Information Visualization for Learning words in the Qur’an", **International Journal on Islamic Applications in Computer Science And Technology**, Vol. 1, Issue 3, December 2013, 75-82, p.82.

[3] F. S. Binti, G. Khattab, C. McKean, "The Qur’an Lexicon Project: A database of lexical statistics and phonotactic probabilities for 19,286 contextually and phonetically transcribed types in Qur’anic Arabic", **Proceedings of the 18th International Congress of Phonetic Sciences.** Glasgow, 2015.

[4] http://corpus.quran.com/wordbyword.jsp

[5] O. Granichin, V. Volkovich, D. Toledano-Kitai, **Randomized Algorithms in Automatic Control and Data Mining.** Heidelberg - New York - Dordrecht - London. 2014. 251p.

[6] D. Shalymov, V. Pavlov, O. Redkin, O.Bernikova, "Arabic manuscripts identification based on Feature Relation Graph", **Proceedings of Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference**, AINL-ISMW FRUCT, 2015, pp. 83-88.