Integration of Language Processing and Linguistic Research as the Mainstream in the Arabic Studies

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

and

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

ABSTRACT

The paper focuses on analysis of dynamics in the various fields of the Arabic language research. It describes new topics that recently have become of high importance for the challenges of the modern society. To this end, we considered papers, published in the scholar database ("Web of Science") in different periods of time. In 2015 speech processing proved to be one of the most popular field of the Arabic research, while in 2005 "pure linguistic" issues (without any mathematical approach) had been more widely spread. The results demonstrate the importance of integration of computer processing into linguistic research. On the current stage, each specialist in humanities should be aware of the computer technologies. Development of the linguistic software and its implementation in the Arabic research influence the language itself. Thus, the spread of ICT made the Arabic language to coin new terminology as well as to create new approaches for the English borrowings adaptation.

Keywords: Arabic, language processing, linguistics, terminology.

1. INTRODUCTION

Invention of digital computational technologies and the Internet may be compared with two informational revolutions in history - development of writing and introduction of printing; all of them completely changed the life of society.

Information and communication technologies (ICT) have become inseparable part of the present-day world. They cover all aspects of life among them the scholar research. ICT provide new opportunities for getting better and quicker results in scientific sphere, they also contribute to formation of new areas of knowledge, especially in the field of interdisciplinary studies. Another special feature of the modern science in the course of its innovative tendencies is focusing on the practical application of results of any research. As a witness of that purely theoretical approach receives less and less support from scientific and educational organizations. Meanwhile humanities in general and linguistics in particular keep up with the times and differ a lot from what they were three or four decades ago. Despite introduction of a huge amount of software, which is commonly used both in

everyday life and education, the linguists have just begun to rely upon methods of computer data processing.

The Arabic language research could be a vivid example of dynamic development of approaches to its processing. The right set of tools configured for Arabic increases productivity and enables research in a community of experts [1]. Being one of the most widely spoken languages on the earth and one of the six languages of the United Nations Organization, it has a lot of linguistic peculiarities which differ a lot from languages, based on the Latin or Cyrillic scripts. Besides that it has vast inflectional and word-formative paradigm, which makes its formalization extremely difficult. In addition to this, the current linguistic situation in the Arab world is characterized by simultaneous usage of Modern Standard Arabic, as a mean of official, primarily written, communication and different local dialects, which are means of oral interaction. This diversity reflects in the methodology of the Arabic language formalization and should be taken into consideration while conducting any studies related to Arabic. Moreover, the usage of Arabic as a mean of e-communications has also contributed development of special linguistic technological solutions as well.

The present research focuses on analysis of dynamics in the fields of the Arabic language research which has happened during the last two decades. It describes new topics that recently became of high importance, considering the challenges of the modern society.

2. MATERIAL AND METHODOLOGY

This study aims at demonstration of the main tendencies in the field of the Arabic language research by means of analysis of the "Web of Science" database content [2]. To this end, we considered papers, published in this database during 2015 in comparison with materials edited in 2005. We searched for publications, which included the keyword "Arabic" in their titles and were limited by two research domains ("Science Technology & Social Sciences"). It is well known that scientific works of humanitarian nature are rarely indexed in the "Web of Science", but the situation changes a lot when speaking about interdisciplinary research. We believe that content of this database reflects the real situation and is appropriate for getting necessary results in terms of current research.

Modern state-of-art solutions for the Arabic language processing influence not only the quality and essence of the linguistic research, but they have an impact on the language itself, primarily on its vocabulary. For the last decades, a big amount of new words denoting computer terminology became widespread. Therefore, in this study we analyze the computer terminology as a new phenomenon in Arabic, and the way of its adaptation.

3. LITERARY REVIEW

Initial stage of the first researches of the Arabic language formalization dates back to 80-90-ies (see, e.g., [3], [4]). Until the beginning of the 21st century, the major topic of the interdisciplinary Arabic language research was automatic processing of its morphology, aimed at finding appropriate solution for stemming. Historiography of development of this field of study could be found in [5].

Later on, spread of the Internet gave a birth to new technologies in the Arabic language processing. At that stage it was necessary to develop algorithms for data retrieving from huge text arrays (see, e.g., [6]), as well as for providing text classification. The only stemming solution became insufficient for solving such kind of tasks and, as a result, it forced to develop Data Mining and its implementation in linguistic research. [7]. Taking into consideration the ever-increasing volume of the Internet content we can conclude

that information retrieval from huge text collections will be an important issue for the years to come [8]. Recent contributions to scientific and technical progress in Arabic computer application have been thoroughly reviewed by Noureddine Chenfour [9]. Meanwhile issues in Arabic computational linguistics were presented by Everhard Ditters [10]. Some aspects of collaboration of humanities and mathematicians in the case of Arabic were discussed in authors' previous paper [11].

4. MODERN TRENDS IN THE ARABIC LANGUAGE RESEARCH: WEB OF SCIENCE CONTENT ANALYSIS

One of the most appropriate sources for analysis of the modern trends in the Arabic language research is the scientific database, which "provides you access to the most reliable, integrated, multidisciplinary research connected through linked content citation metrics from multiple sources within a single interface" [2]. Web of Science seems to be one of the most popular sources of this kind, that is why we have chosen it for the purposes of our investigation. The search conditions were as following:

- *Search word*: Arabic (in: Title).
- *Publication year:* 2015.
- *Research domain:* Science Technology & Social Sciences.
- *Research areas:* Linguistics, Computer Science.

As a result, we received 62 publications of different topics; the most of them are of interdisciplinary nature, i.e. aimed at developing of special computer solutions for getting results, based on linguistic data (more, than 80%). After that we carried out similar experiment but for the year 2005 and came to a strong conclusion that "pure linguistic" issues (without any mathematical approach) had been more widely spread just a decade ago, that differs a lot from what we witness today.

Meanwhile classification of the publications of 2015 gave us another result concerning interdisciplinary studies in the area of dialect processing (see, e.g., [8]) vs. Modern Standard Arabic. The ratio in this case is approximately 20% vs. 80%.

Classification of the publications in accordance with their topics is summarized in the following Table. Results include topics, which were presented in more than two publications (out of 62).

Table 1. Thematic classification of scholar publications, indexed in the Web of Science.

Dialect processing / speech processing	Handwriting processing / manuscript	Event retrieval / Author Identification	Stemming	Teaching of Arabic	OCR
28%	16%	14%	12%	10 %	6%

Results show that new task, i.e. speech processing has become popular. At the same time OCR and stemming receives less attention. We believe that the results, based on the Web of Science reflect the real situation and demonstrates modern trends in the Arabic language processing.

5. COMPUTER TERMINOLOGY IN ARABIC

Since the 'birth place' of the most of the computer software in the 80-s 90-s of the last century was the USA, the "computer language" was English. Therefore on the initial stage the regional localization of software products, i.e. adaptation for the users in the countries of the Middle East along with the alignment of text from right to left also presupposed translation of computer terminology from English into Arabic and its substitution with direct equivalents.
For instance:

"Windows" -
النوافذ /an-nawāfiḏ/,
"search" -
بحث /baḥṯ/,
"*view*"-
عرض /ʻarḍ/ etc.

Later on exponential development of the software which had been designed for various spheres of life took place as well as number of users in the rest of the world who were non-English speakers greatly increased. The new reality forced computer specialists to take into consideration cultural and language peculiarities of the targeted audience and to make precise choice of synonyms or local variants of words or combinations of words rather than rely on the direct translation of terms.

Integration of Language Processing and Linguistic Research may be regarded not only in the context of development of the Arabic language processing but it influences the structure of the language itself. Development of ICT and its implementation in everyday life created a lot of new words for connote the so-called computer terminology. It is believed that most words denoting computer terms, are borrowed from the English language. Without English technical terminology people would not be able to use computers. [9] Nevertheless the Arabic language demonstrates its peculiarities in technical borrowings adaptation. First of all it is essential for the Arabic language to extend the use of original Arabic roots and well as descriptive forms or direct literal translation. For example:
"machine translation" –
ترجمة الية /tarğama āliyya/
"*digital*" –
رقمي /raqmī/
"*file*"-
ملف /milaff/.

The tendency to use Arabic roots for new terminology was initiated by the Academies of the Arabic language [13]. In order to analyze the functioning of a computer terminology in Arabic we examined the Microsoft glossary, which includes more than 12 000 words and expressions [14], as well as the terms found in various Arabic-speaking Internet sites dedicated to computer terminology. The results show that the approximate percentage of the direct borrowings in computer terminology in Arabic is just 4 %:
"*web*" –
ويب /wēb/
"*video*" –
فيديو /fīdīū/
"*protocol*" –
بروتوكول /brūtūkūl/
Frequently such kind of borrowings retain Latin alphabet:
"OCR menu" – قائمة OCR

Moreover, the names of major brands and companies (such as Microsoft, Google, Microsoft Windows, Microsoft Word, etc.) are not translated into Arabic and are not transliterated by means of Arabic script.

Another peculiarity of computer terminology in Arabic is variation in translation of the same terms in different sources. As an example, we give the translation of the term "*Firewall*" (part of the operating system, which protects the computer against illegal network attacks). In Arabic, there are two options for translation of this term: one is a calque of the English words combination - جدار النار (which literally means "*a wall of fire*"). The second is a descriptive version - جدار الحماية (*"protection wall").*

Even though there are indigenous Arabic analogues for the computer terminology, currently there is a growing tendency of usage of their English equivalents. For example the word "computer" may be translated as حاسوب /ḥāsūb/ and كمبيوتر /kumbyūter/. The first word was coined by the Academies of the Arabic language and is associated with the real Arabic word, nevertheless the second is more commonly spread. At the same time spread of English increased significantly with the escalation of implementation of computers in everyday life. "When the need for global

communication came to exceed the limits set by language barriers, the spread of English accelerated" [9].

6. CONCLUSIONS

Modern trends in the development of science suggest that the use of interdisciplinary approaches provide successful development of any field of knowledge. This is especially true for the humanities, which recently mostly relied on the results of empirical research and scholar's perception and evaluation of phenomenon or object. In this case the results may vary depending not only on the researcher's knowledge, his experience, methodology and intuition but also on his attitudes, stances and sympathies, and even likes and dislikes as well. The solid data provided by computer minimizes such influences and narrows the area for misinterpretations of results.

Thus, integration of language processing into linguistic research became essential part of the Arabic studies. In the present research, we defined modern trends in the interdisciplinary approaches to the Arabic language research. Results include new topics, which recently became of high importance for the needs of modern society. Despite the developments of new research attitudes and technologies there is a number of problems related to Arabic text processing which still remain far from their final solution. It seems to be more effective to combine traditional methods of investigation based on segmentation along with newly developed methods of text processing considering text as an indivisible object with unique combinations of interchanging relevant markers.

The next task is to define these groups of markers and to develop mathematical algorithms for their analysis and processing. In this case, most of the existing approaches in textual processing must be revisited and re-examined likewise new models of mathematical analysis should be created in order to overcome the limitations of the previous techniques.

Development of the linguistic software and its implementation in the Arabic research influence the language itself. Thus, ICT evolution made the Arabic language to coin new terminology as well as create approaches for the English borrowings adaptation. The latter requires from a scientist to be competent and aware of different fields of study.

Multifaceted scholar activities of a researcher are considered to be the hallmark of the ancient times and the Middle Ages. It seems that the history repeats and now we are witnessing reflection of the past but on a new stage of social development.

7. ACKNOWLEDGMENTS

The authors acknowledge Saint-Petersburg State University for support this research.

8. REFERENCES

[1] W. Samy, Internet, **Encyclopedia of Arabic Language and Linguistics**, Managing Editors Online Edition: Lutz Edzard, Rudolf de Jong. Consulted online on 18 February 2017 <http://proxy.library.spbu.ru:2083/10.1163/1570-6699_eall_EALL_COM_vol2_0052>
First published online: 2011
First print edition: ISBN: 9789004177024, 20090831

[2] **Web of Science™. Thompson Reuters**. Web: https://apps.webofknowledge.com/UA_GeneralSearch_input.do?product=UA&search_mode=GeneralSearch&SID=W1FotyYKuJym4O9Aa9w&preferencesSaved=. Retrieved: August, September 2016.

[3] P. Toma, SYSTRAN as a multilingual machine translation system-archive. **Overcoming the language barrier**, München: Verlag Dokumentation, 3-6 May 1977, Vol. 1, pp: 569–581.

[4] J. McCarthy, A prosodic theory of nonconcatenative morphology. **Linguistic Inquiry**, 12, 1981, pp. 373-418.

[5] M. Dahab, A. Al-Ibrahim, R. Al-Mutawa, A Comparative Study on Arabic Stemmers. **International Journal of Computer Application**, 2015, 125(8), pp. 38-47.

[6] O. Bernikova, O. Redkin, The Arabic Language Processing: Peculiarities of Interdisciplinary Research. **Proceedings of the 9th International Multi-Conference on Society, Cybernetics and Informatics**, Orlando, USA, 2015, pp: 81-86.

[7] O. Granichin, V. Volkovich, D. Toledano-Kitai, **Randomized Algorithms in Automatic Control and Data Mining**. Springer-Verlag: Heidelberg, New York, Dordrecht, London. 2015.

[8] Al-Ghuribi, S. Alshomrani, Bi-languages Mining Algorithm for Extraction Useful Web Contents (BILEx). **Arabian Journal for Science and Engineering,** 40(2), 2015, pp. 501-518.

[9] N. Chenfour, Automatic Language Processing, **Encyclopedia of Arabic Language and Linguistics**, Managing Editors Online Edition: Lutz Edzard, Rudolf de Jong. Consulted online on 19 February 2017 <http://proxy.library.spbu.ru:2083/10.1163/1570-6699_eall_EALL_COM_0031>
First published online: 2011.

[10] E. Ditters, Computational Linguistics, **Encyclopedia of Arabic Language and Linguistics**, Managing Editors Online Edition: Lutz Edzard, Rudolf de Jong. Consulted online on 18 February 2017 <http://proxy.library.spbu.ru:2083/10.1163/1570-6699_eall_EALL_COM_0064>
First published online: 2011
First print edition: ISBN: 9789004177024, 20090831

[11] O. Bernikova, O. Redkin, Humanities and Mathematical Approaches in the Case of Arabic, **Proceedings of the 20th World Multi-Conference on Systemics, Cybernetics and Informatics**, Orlando, USA, 2016, pp: 279-283.

[12] I. Youssef, Vocalic Labialization in Baghdadi Arabic: Representation and Computation. **Lingua,** 160, 2015, pp. 74-90. DOI: 10.1016/j.lingua.2015.04.001.

[13] A. Atawneh, English Loanwords, **Encyclopedia of Arabic Language and Linguistics,** Managing Editors Online Edition: Lutz Edzard, Rudolf de Jong. Consulted online on 19 February 2017 <http://proxy.library.spbu.ru:2083/10.1163/1570-6699_eall_EALL_COM_vol2_0006>
First published online: 2011
First print edition: ISBN: 9789004177024, 20090831

[14] **Microsoft Language Portal.** Web: https://www.microsoft.com/Language/en-US/Default.aspx. Retrieved: December 2016.