(English translation)

# REVIEW

on Competition procedure for the position of

**Professor**

in the professional field 4.6 Informatics and Computer Science, specialty Informatics

Announced by Institute of Information and Communication Technologies – Bulgarian Academy of Science, for Section "Artificial Intelligence and Language Technologies" in the State Gazette, issue 45/28.05.2021 and on the Institute of Information and Communication Technologies – Bulgarian Academy of Science web page

**By Prof. Dr. Tinko Velichkov Tinchev,**

Sofia University St. Kliment Ohridski, Faculty of Mathematics and Informatics, professional field 4.5 Mathematics (Mathematical Logic) in his capacity of Scientific Jury Member

following

Order# 185/ 27.07.2021 of the Director of the Institute of Information and Communication Technologies – Bulgarian Academy of Science

The only candidate in this competition procedure for Professor is **Stoyan Milkov Mihov, D.Sc., currently Associate Professor** at the Institute of Information and Communication Technologies – Bulgarian Academy of Science (IICT-BAS), Section "Artificial Intelligence and Language Technologies"

## I. GENERAL DESCRIPTION OF THE MATERIALS PRESENTED

### 1. Brief biographical data

Assoc. Prof. Stoyan Milkov Mihov was a student at the Faculty of Mathematics and Informatics (FMI) at Sofia University "St. Kliment Ohridski" from 1988 to 1993 and graduated with a master's degree in mathematics with a diploma thesis in mathematical logic

on the topic "Unification of co-regular sets". Since 2000 he has been Doctor in Informatics with a dissertation "Minimal acyclic automata: constructions, algorithms, applications". Since 2020 he has been Doctor of Science in the professional field 4.6 Informatics and Computer Science with dissertation "Finite-State Automata, Transducers and Bimachines: Algorithmic Constructions and Implementations".

Since 1995 and until now Assoc. Prof. Stoyan Mihov has been working at IICT-BAS (including CCIIT, CLPOI and IPOI, of which IICT is the legal successor) consecutively as a programmer, assistant, chief assistant, and since 2006 - an associate professor in the section "Artificial Intelligence and Language Technologies". He is an established scientist in the international scientific community with one monograph and over 60 publications in renown specialized scientific journals and papers at highly valued scientific conferences, 33 of which are in SCOPUS and / or WoS and which have over 369 SCOPUS citations (without self-citations) and h-index 7 (without self-citation).

Assoc. Prof. Mihov has solid experience in fundamental and applied research with industrial applications. He was national coordinator, leader and participant in a number of European, international and national projects (IMPACT, OCoRrect, AComIn, CULTURA, eHealth in Bulgaria (e-health), SpeechLab and etc.). Additionally, Assoc. Prof. Mihov was a senior researcher for companies Commek EOOD and Rila Solutions EAD. He holds a patent in USPTO, active since 2011.

Assoc. Prof. Stoyan Mihov has been active in teaching since 2003 in the bachelor's programs of FMI, with particularly valuable contributions in the master's programs Logic and Algorithms (specializaion Mathematics and specializaion Computer Science) and Computer Linguistics (specializaion Computer Science) at the Department of Mathematical Logic.

He is the supervisor of 10 successfully defended master's theses, 8 in FMI and 2 in the Faculty of Slavic Philology. Under his supervision, 2 doctoral dissertations were successfully defended (1 in FMI and 1 in IICT).

## 2. General description of the submitted documents

The documents submitted by Assoc. Prof. Stoyan Mihov for participation in the competition are in full compliance with the requirements of the Law on Development of Academic Staff in Republic of Bulgaria (ZRASRB), the State Rules of Procedure for its application, the Rules for Acquisition of Scientific Degrees and Occupation of Academic Positions and with the specific requirements in the regulations of BAS and IICT-BAS. They include:

- application to the Director of IICT for admission to the competition;
- CV - European format;

• a copy of the diploma for educational and scientific degree "Doctor";

• a copy of the diploma for the scientific degree "Doctor of Science";

• IICT certificate with outgoing number 395 from 24.06.2021 for work experience (28 years, of which 15 years and 1 month as an "associate professor");

• list of publications for participation in the competition (18 issues, not used in previous procedures);

• list of citations from Scopus (213 pieces not used in previous procedures);

• summaries of the publications for participation in the competition (in English and in Bulgarian);

• copies of the publications for the competition (18 copies, not used in previous procedures);

• statement for fulfillment of the minimum national requirements under art. 2b, para. 2 and 3, and the requirements of IICT under Art. 2b, para. 5;

• list of the original scientific and scientific-applied contributions of the publications presented for the competition;

• a declaration that there is no legally proven plagiarism in the publications;

• a declaration of co-authorship by Prof. Klaus Schultz of LMU Munich;

• an email from the technical editor Bianca Truthe of the Journal of Automata, Languages, and Combinatorics, certifying the acceptance for publication of the article number 4 from the list of publications for participation in the competition;

**3. General characteristics of the publications for participation in the competition**

For participation in the competition, Assoc. Prof. Stoyan Mihov presents 18 publications thematically very well focused in 3 areas - 16 articles, one preprint from Arxiv and one patent in the US Patent Agency active since 2011. All 16 articles are contained in Scopus and / or WoS. They carry the following points for indicator G7 (according to the Minutes of the 48th meeting of the Seventh General Assembly of BAS, the coefficient for G7 is 2):

• the articles [1] and [2] are respectively in the journals Natural Language Engineering (2007) and Theoretical Computer Science (2019), which for these years are in quartile Q4 for WoS, due to which each of them carries 24 points;

• the articles [3], [4], [5], [6] and [8] are in editions referenced by Scopus and are with SJR, therefore each of them carries 20 points;

• the article [9] is in the journal Computational Linguistics (2006), which is in quartile Q1 for 2006, so it carries 50 points;

• The articles [10], [11], [13] and [15] are in editions referenced by Scopus and are with SJR, therefore each of them carries 20 points.

Therefore, the points for indicator G7 are 2x24 + 5x20 + 1x50 + 4x20 = 278. Since the patent [18] carries 25 points for indicator G9, the points for indicator G of the national minimum criteria calculated according to the decision of the "Minutes of the 48th meeting of the Seventh General Assembly of BAS" are 303, which means that the requirement of the specific rules of IICT for at least 260 points according to indicator G has been met.

The attached list of citations of articles available in SCOPUS by Assoc. Prof. Stoyan Mihov from articles that are also in SCOPUS contains 213 items. This list does not contain self-citations and those already used in previous procedures. According to the decision of the "Minutes of the 48th meeting of the Seventh General Assembly of BAS" the points on indicator D are **619**, significantly more than the required by the specific requirements of IICT 140 points.

All submitted publications are joint. I accept the equal contribution of the authors, given the declared by Assoc. Prof. Stoyan Mihov and the declared by Prof. Klaus Schultz. In fact, Prof. Schultz's declaration is in the context of a well-motivated recommendation for the election of Assoc. Prof. Stoyan Mihov as professor.

**4. Analysis of the scientific and scientific-applied achievements of the candidate in the publications for participation in the competition**

The research of Assoc. Prof. Stoyan Mihov is largely stimulated by its applicability to non-trivial practical problems for natural language processing, correction and normalization of texts, speech recognition and generation. I would like to note that the mentioned specific applications, regardless of their independent value, are evidence of a meaningful richness of the abstract structures and operations on them, studied by Assoc. Prof. Mihov. Of course, the relationship between abstract models and practical language applications is usually not linear - different specific large data collections, corpora, are needed. The development of a methodology for the creation, the creation of corpora and their evaluation requires not only a lot of resources, but also a special kind of creativity inherent in scientific and applied activities. In the 18 publications submitted for the competition there are a number of significant scientific and scientific-applied contributions, which are correctly reflected in the author's reference and where an appropriate grouping of the publications in three necessarily intersecting directions of the contributions.

A. **Contributions in finite automata theory.** Articles [1] - [4] from the list of submitted publications belong to this group. Articles [1] and [4] consider two variants of the text

rewriting task: given final rewriting dictionary for (list of pairs of strings/words) a subsequential transducer is built, which for a given text outputs the intended rewriting result. In the first case, the substitution is context-free, while in the second case, left and right contexts are allowed, with the left context being set with an arbitrary regular expression, and the right context being a finite list of strings/words. The constraint on the right context is essential for the existence of a subsequential transducer. In both cases, substitution conflicts are resolved according to the leftmost-longest match strategy: substituting the first from left longest match with the corresponding string/word. The presented efficient construction of a subsequential transducer using fail transitions and an algorithm for composition of such kind of transducers are significant scientific contributions.

In the article [3] efficient algorithms for composing a conditional probabilistic subsequential transducers with a probabilistic failure subsequential transducers are developed. This is a significant scientific contribution that has applications in speech recognition.

The article [2] develops building of bimachines relying on new principle. In particular, the construction of an almost optimal bimachine from a given functional real-time transducer is defined, which is a significant scientific contribution.

**B. Contributions to natural language processing and speech recognition.** This group includes publications [5] - [8] and [18].

The article [5] presents the first implemented system for speech recognition of continuous Bulgarian speech in a large dictionary, which is based on hidden Markov models using Gaussian mixtures. A significant improvement of the system is presented in the article [6], due to the construction of a suitable deterministic finite automata, efficiently representing a wider range of candidates than those used in [5]. The statistical significance of the improvement is shown. This algorithm which re-evaluate candidates lattice is a significant scientific contribution. The implementation of these systems is an indisputable scientific-applied contribution.

The article [7] presents a created Bulgarian speech corpus with records of 147 speakers, containing 21891 utterances with a total duration of about 32 hours. This corpus, used to train the system by [5], significantly improves the recognition of legal texts after adaptation to a speaker. It turns out that this corpus is insufficient for technologies based on deep neural networks. The article [8] presents a corpus containing records of 572 speakers with a total duration of about 250 hours. Various natural language processing techniques and speech recognition technologies have been used to create this corpus. The creation of these two corpora, based on different technologies, as well as the methodology, are scientific-applied contributions.

The publication [18] is an active patent in the US Patent Office. It is a methodology for analyzing the relationships between different subjects, based on the specific information retrieved from large arrays of texts in natural language with subsequent analysis of a properly constructed graph. The tool built on this methodology is useful for measuring, analyzing and planning media communication and is a nice usefull scientific-applied contribution.

**C. Contributions in the field of approximate search, text correction and normalization of texts.** In this group are the publications [9] - [19]. In the author's reference of Assoc. Prof. Mihov they are appropriately divided into 3 subgroups.

C1. *Approximate search algorithms*, [13] - [15]. The article [13] is an invited lecture in the special session Computational Linguistics at the 7th Conference on Computability in Europe, CiE 2011, Models of Computation in Context. In it is discussed the view of the similarity search as a special kind of computation, which is an undoubted scientific contribution with a methodological nature. An original approach to appriximate search is the algorithm presented in [14], aimed at avoiding the so-called wall effect. The algorithm showed particular efficiency in huge dictionaries of very long words (records) and long distances. A specific implementation of it is presented in [15]. With this implementation, the author's team won the String Similarity Search competition, organized within the joint conference EDBT / ICDT 2013. This algorithm, called WallBreaker, is a pleasant significant scientific contribution.

C2. *Text correction*, [9] - [12]. The article [9] presents a methodology for automatic and semi-automatic compilation of dictionaries with incorrect spellings of words and a methodology for automatic creation of dictionaries with correctly spelled words on web pages. A methodology for orthographic correction has also been created, allowing a significant improvement in the quality of the corpora automatically retrieved from websites. These methodologies were used in the articles [10] and [11] to create a universal Levenstein automaton for finding the lists of candidates for corrections. The algorithm in [10] uses a manually corrected corpus of noisy words, while the algorithm in [11] gives similar results, but without the use of such a dictionary. The article [12] is their subsequent development, in which the list of candidates for correction is ranked by means of bigram language model.

C3. *Normalization and modernization of historical texts*, [16] and [17]. In the article [16] with the help of infix structures based on finite state automata, a very good correction of historical variations of words is achieved, using a manually created corpus of observed spelling variations. In [17] the ideas from [16] were continued with the use of the so-called functional automata, as the problem of text normalization is

interpreted as a special case of the machine translation problem. The results of the comparative analysis with the well-known specialized systems VARD2 and Moses show significantly higher quality of the developed algorithm.

**5. Characteristics and evaluation of the teaching activity of the candidate**

Assoc. Prof. Stoyan Mihov has significant experience and high results in teaching at Sofia University "St. Kliment Ohridsky" with lectures at the Faculty of Slavic Philology (FSLF) and at Faculty of Mathematics and Informatics (FMI). In FSLF he taught the course Formal Languages and Grammars in the master's program "Computerized Humanities" in the period 2003-2006. Since 2015 he has been teaching the course "Search and Retrieval of Information. Deep Machine Learning", which is essential for the Master's program in Computational Linguistics, but is open to other master's and bachelor's programs. This is an in-depth course on contemporary approaches to the topic.

A special place in the teaching activity of Assoc. Prof. Mihov is occupied by the course "Applications of finate authomata" for master's programs at the Department of Mathematical Logic, open for bachelor's programs at FMI, which he has been giving since 2003. This course adds specific theoretical and applied color to the master's program "Logic and Algorithms". It, along with the classic results for finite trasducers, contains a constantly updated part with the latest results in the field and practical application of algorithms.

It enjoys great interest among the best students and is a source of topics for master's thesises. Under the supervision of Assoc. Prof. Stoyan Mihov, 8 diploma theses and 2 doctoral dissertations (one in FMI and one in IICT) have been successfully defended and, currently, one doctoral student is in the process of training at IICT.

**6. Critical remarks and recommendations**

I have no significant critical remarks.

**7. Personal impressions of the candidate**

I have personally known Assoc. Prof. Stoyan Mihov since 1992, when he was one of the best students in the master's program at the Department of Mathematical Logic. From the seminar, which Assoc. Prof. Mihov has been leading for twenty years at IICT, and his participation in seminars and conferences related to the Department of Mathematical Logic at FMI, I have witnessed his professional development. I highly appreciate his teaching activity at FMI. Over the years we have had fruitful joint work on 3 projects.

Along with his impressive erudition in a number of areas of knowledge and creative activity, I would like to note his ability to organize and lead teamwork. He has the rare ability to see an abstract form of a practical problem, to solve this abstract form and to implement the solution in a complete technological industrial product.

**8. Conclusion on the application**

After getting acquainted with the materials and publications presented in the competition, based on the analysis of their significance and the impressive scientific and scientific-applied contributions contained in them, I confirm that the scientific and scientific-applied achievements meet the requirements of the Law on Development of Academic Staff in Republic of Bulgaria (ZRASRB), the State Rules of Procedure for its application, and the respective Regulations of BAS and IICT-BAS for holding by the candidate the academic position "Professor" in the scientific field of the competition.

In particular, the candidate Assoc. Prof. Stoyan Milkov Mihov satisfies the requirements of IICT-BAS (and therefore the minimum national requirements) in the professional field and no plagiarism has been established in the publications submitted under the competition.

I give my positive assessment of the candidacy.

**II. OVERALL CONCLUSION**

Based on the above, I strongly recommend to the scientific jury to propose to the Scientific Council of IICT-BAS to elect Assoc. Prof. Stoyan Milkov Mihov, D.SC. to take the academic position of "Professor" in the professional field 4.6. Informatics and Computer Science, specialty Informatics.

September 23, 2021



Prof. Tinko Tinchev, PhD