OPINION
for the competition for the academic position "Professor"
in the Professional Field 4.6 Informatics and Computer Science
Scientific Specialty Informatics
for the needs of Department of Linguistic Modelling and Knowledge Processing
Institute of Information and Communication Technologies (IICT)
Bulgarian Academy of Sciences,
announced in Newspaper of State, No. 45 of 28.05.2021

The opinion is written by **D.Sci. Vesselin Stoyanov Drensky, Full Member of the Bulgarian Academy of Sciences**, retired professor in the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences as a member of the Scientific Jury for the competition by Order No. 185 of 27.07.2021 of the Director of IICT.
The only applicant who has applied for the position is D.Sci. Stoyan Milkov Mihov, Associate Professor in IICT-BAS.

For his participation in the competition the applicant has submitted all required by the law and the accompanying rules of IICT. In addition, the documentation contains a recommendation (in German and English) from Prof. Klaus Schulz from the University of Munich, who is one of the main coauthors of the applicant. Although this is not required by the Law and the Regulations, it would be useful to attach to the documents an excerpt from the competition announcement in the Newspaper of State and a list of all publications of the applicant, which would give a more complete picture of his achievements. (There is such a list on the IICT website, but it has not been updated since 2019.)

1. **Biographical data.** Assoc. Prof. D.Sci. Stoyan Mihov graduated with a Master's degree in Mathematics from the Faculty of Mathematics and Informatics at Sofia University "St. Kliment Ohridski" (with Master Thesis "Unification of Regular Sets" and supervisor Assoc. Prof. Dr. Anatoly Buda). He defended his Ph.D. Thesis in Informatics at IICT on "Minimal Acyclic Automata: Constructions, Algorithms, Applications" with supervisor Prof. D.Sci. Dimitar Skordev. Again in IICT he defended his Doctor of Science Thesis in Informatics and Computer Science on the topic "Finite-State Automata, Transducers and Bimachines: Algorithmic Constructions and Implementations". The entire scientific career of Assoc. Prof. D.Sci. Mihov is connected with IICT and its predecessors. He has been a programmer, assistant professor, chief assistant, and since 2006 he has been an associate professor. (In the CVs in the IICT website and in the documents submitted for participation in the competition there is a certain discrepancy in the years when the applicant started his career at BAS.) Since 2003 he has been a part-time lecturer at the Faculty of Mathematics and Informatics at Sofia University. In addition, he was a head of a group at Rila Solutions and the head of the Department of Research Activities (with an additional employment contract) at COMMEQ.

2. **General characteristic of the scientific work and achievements of the applicant.** The main directions of research of Assoc. Prof. D.Sci. Stoyan Mihov are related to theoretical informatics (theory of finite automata - at the meeting point with the abstract theory of computability, approximate search, synthesis and recognition of speech, computational linguistics) and practical implementation of many of the results with specific applications in natural language word processing, speech recognition, text correction and normalization. Most of the created techniques are theoretically effective and are applied to problems which are important from a practical point of view. The applicant is an author (in

most of the cases a coauthor) of more than 60 scientific publications, including one monograph in a recognized international publishing house. More than half of the papers are with impact factor or SJR. The publications are for the period from 1989 until now. According to the data presented in the CV of the applicant, his publications are cited more than 420 times in more than 340 documents. A list of 213 citations is given in the attached document for covering of the minimum requirements of the law and the rules of the IICT, as one of the articles (with Schulz from 2003) is cited 93 times. In particular, I want to emphasize that the applicant is the author of the programming language C(M), which directly converts mathematical constructions into programs in the programming language C. This allows rapid computer implementations of complex algorithms and has a wide range of applications.

3. **Characteristics and evaluation of the teaching activity and the participation in projects of the applicant.** Assoc. Prof. D.Sci. Stoyan Mihov has a long and successful teaching career. He has red and continues to read a number of courses at FMI and the Faculty of Slavic Philology (FSPh) at Sofia University "St. Kliment Ohridski". He has two successfully defended Ph.D. students (in IICT and FMI at Sofia University), 8 defended graduates in FMI and 2 in FSPh at Sofia University. To these data I shall add the extremely positive opinion of Prof. Schultz on the pedagogical activity of the candidate. According to Prof. Schulz, in his lectures and seminars Assoc. Prof. D.Sci. Stoyan Mihov includes many interesting and deep scientific topics, which motivates students. As a research supervisor, he offers his Ph.D. and Master's students problems that lead to interesting results and direct practical use in applied projects. Since 1993, Assoc. Prof. D.Sci. Stoyan Mihov has been actively involved in a number of scientific and applied projects, in many of which he applies his developments in practice. Since 1996 he has been the principal investigator of most of the projects in which he participates. He is currently the head of a work package in the National Research Program "Electronic Health Care in Bulgaria".

4. **General description of the submitted materials.** The applicant has submitted for participation in the competition 15 scientific publications published in the period 2006-2021, 1 publication accepted for publication, 1 preprint in the popular preprint database arXiv.org and one patent registered in the USA. Of the articles, 4 are in journals (in Natural Language Engineering, Theoretical Computer Science, J. Automata, Languages and Combinatorics, Computational Linguistics), 5 are in series (Lect. Notes Comp. Sci - 4, Studies in Comp. Intelligence - 1) and 7 are in conference proceedings. All 16 publications published or accepted for publication are in issues with SJR. The applicant has not stated this, but I think that two of the journals have an impact factor in the year of publication. It would be useful if the applicant had also provided data on the exact SJR of the journals, series and proceedings. All papers are with coauthors (4 with one coauthor, 8 with two coauthors, 3 with three, one with five and one with six coauthors). Among the coauthors are students and colleagues of the applicant, 4 are scientists from Germany and 1 is a scientist from Canada. Particularly impressive is the collaboration with Schulz (coauthor of 10 of the articles presented in the competition) and with the successfully defended under the supervision of the applicant PhD students Mitankin and Gerdjikov (coauthors of 8 and 5 articles, respectively). The applicant has declared that in all publications the coauthors have equal participation. I believe that in this case the joint papers are a positive fact, because the research is interdisciplinary and at the border on several areas. Collaboration increases the efficiency of research, allows the use of methods from different fields of science and shows the ability to work in a team. Personally, I appreciate the qualities of successful teamwork and research on the meeting point of several areas. According to the information I have, the publications presented have not previously been used for other procedures. For the

competition the applicant has submitted a list of 213 citations of 6 of his articles; 23 of them are for a paper submitted for participation in the competition. The presented table shows that the candidate completely satisfies, and in some of the criteria significantly exceeds the minimum requirements of IICT for participation in a competition for "professor". I did not find any plagiarism in the works submitted for the competition.

5. **Main scientific and scientific-applied achievements.** I shall briefly discuss the main results contained in the submitted works of the applicant, as well as my assessment of them. In the description of his scientific contributions the applicant has divided his publications into three groups: (1) Theory of finite-state automata (papers [1-4] from the list of publications for participation in the competition); (2) Natural language processing and speech recognition (papers [5-8] and patent [18]); (3) Approximate search, correction and normalization of texts (papers [9-17]). As the candidate notes, this division is conditional because many of the results fall in more than one direction.

(1) **Theory of finite-state automata.** The theory of finite-state automata and the related with it theory of formal languages are important branches of the contemporary mathematics and theoretical computer science with numerous applications in other branches of mathematics (e.g. in algebra, mathematical logic and combinatorics) as well as in other branches of knowledge including applications for the solution of practical problems. In particular, these theories are largely applied in linguistics in the solutions of difficult problems in text and natural language processing, speech processing, pattern matching, approximate dictionary search, text correction, etc. Paper [1] of 2007 deals with the task of rewriting text in a dictionary-based text rewriting. The authors propose an effective method for constructing a subsequential transducer, which for a given input text gives the output text for a time that depends linearly on the length of the input and output texts. The method has two advantages - optimal efficiency and standard possibilities for combining with other transducers to solve more complex problems in just one step. The other three papers are from the last two years. In paper [4] the construction in [1] is improved, which expands the possibilities for application, further improves the efficiency and allows work with very large dictionaries. The paper includes data for an experiment with the construction of a transducer which replaces the Wikipedia keywords with a link to the corresponding page. Paper [2] deals with a new construction of bimachines and is a further development of a previous work which is not included in the competition. Bimachines form a class of deterministic finite-state machines which present the class of regular functions on words. There is a standard construction of bimachines starting with functional transducers. In previous studies, the applicant had a new construction which transfers the transducer directly into a bimachine. Now this construction has been improved, and a number of difficulties, including algebraic ones, had to be overcome. In particular, a class of monoids is introduced for the needs of the construction, which includes free monoids, groups and other algebraic objects. The new construction is close to optimal. An original methodology for constructing probabilistic converters is proposed in [3]. The advantage of the method is that in the process of work the probabilistic values are correctly preserved, and in case of failure in the execution of any step the determinism of the transducer is preserved. Specific constructions and experimental results in speech recognition are presented.

(2) **Natural language processing and speech recognition.** Papers [5, 7, 8] are devoted to specific implementations of natural language processing systems and speech recognition. The developments are the result of the work of a team led by Assoc. Prof. D.Sci. Stoyan Mihov. Paper [5] describes the implementation of the first system for recognition of continuous Bulgarian speech in a large dictionary for the purposes of working with legal documents. Experiments show that errors are <12% for legal texts and <16% for general

texts. Paper [7] describes the principles used for the construction of the BulPhonC speech corpus. Recordings of 147 speakers with about 22,000 expressions with a total duration of about 32 hours were used to make the corpus. Applied to the system of [5], the development gives an error below 7% at the word level for legal texts. It has turned out that the BulPhonC corpus is too small for the needs of machine learning of deep neural networks. In [8] the corpus of speech BG-PARLAMA is presented, which overcomes this problem. The records of the corpus were taken from the plenary sessions of the National Assembly for a period of 10 years. The amount of data used is impressive - 250 hours of recordings of 572 speakers. We shall pay attention that the used technologies give the advantage to work almost completely automatically with minimal human resources. Paper [6] is of a completely different, more theoretical nature. The theory of deterministic finite-state automata is used. As a result, in addition to finding the few best candidates for correct transcription based on the experiments performed, a significantly larger number of candidates are given without the use of additional time. The experiments performed show less than 4% errors. The patent [18] registers a methodology for analyzing the influence of individual subjects in a given media event. Graph theory is used, applying a variety of graph analysis techniques. This leads to practical methods for evaluating and improving the effectiveness of media communication.

(3) **Approximate search, correction and normalization of texts**. The papers from this group were published in the period 2006-2014, and the second version of the preprint [14] is from 2015. The research is in three subfields: approximate search algorithms (papers [13-15]), text correction (papers [9-12]) and normalization and modernization of historical texts (papers [16-17]).

3.1. Approximate search algorithms. Paper [13] presents applications in natural language processing, and the approximate search is considered as a special type of calculation. Paper [14] presents a new effective method for approximate search in an electronic dictionary, achieving efficiency better than that of existing techniques and offering interesting possibilities for searching in collections of long strings (e.g. sentences). Paper [15] describes the WallBreaker approximate search system with which the team participated in a competition organized by the Humboldt University in Berlin. A number of fundamental and technical difficulties have been overcome. For example, it was necessary to create a new presentation of the dictionary embedded in the system.

3.2. Text correction. Paper [9] analyzes the orthographic (spelling) errors and their correction in Internet and produces statistics of errors of different types. A methodology for compiling dictionaries with incorrect spellings and correctly spelled words is presented. The methods can be used in corpus linguistics. In papers [10-11] a variant of the Levenshtein metric is given, which can be effectively realized by a universal Levenshtein automaton. While the methods in paper [10] require a corpus of orthographic errors and their manually corrected copies, in [11] this problem is solved by creating a methodology for automatical creating of dictionaries with misspellings. Paper [9] shows that the accuracy of the correction is improved by adding the frequency of the bigrams of the words from the Internet corpora as a new factor for assessing the aproximation and the uniform frequency of the words. Experiments show that language models from Internet corpora give better results than those obtained from standard corpora.

3.3. Normalization and modernization of historical texts. In papers [16-17] within the REBELS project a new approach was developed, which is a continuation of the original methodology for normalization and modernization of historical texts, created in the CULTURA project. Techniques from the theory of finite-state automata are used again. Experimental results are presented, which are better than those obtained with existing systems for normalization of English texts and for machine translation.

In conclusion of my comments on the scientific and scientific-applied contributions of the applicant I shall note that the applicant understands very well the main problems in the field. He knows in detail the existing theoretical and practical tools for their solution and uses a rich arsenal of methods. The reliability of the arguments and the results of the experiments is not in doubt. I have not noticed any significant inaccuracies.
The description of the contributions of the applicant and the abstracts of the submitted documents correctly reflect the main contributions of the works submitted for participation in the competition.

6. **Significance of contributions to science and practice.** The results obtained in the scientific and applied work of the applicant are interesting and meaningful. The results and the created methods can be used directly, and some of them are already used successfully in computational linguistics.

7. **Critical remarks and recommendations.** I have already mentioned some suggestions regarding the preparation of the competition documentation. In addition, I would recommend the candidate to publish in even more respectable journals, especially from a formal point of view. This would increase both its international prestige and would bring immediate benefits (scientific and financial) for IICT and for BAS in general. I am convinced that the submitted materials are completely sufficient to meet the formal and scientific criteria for the academic position of professor, but I believe that, as a doctor of science, the candidate should present a more representative (in terms of citations and formal metrics) collection of his results. I consider my remarks to be well-meaning and friendly.

8. **Personal impressions for the applicant.** I know Assoc. Prof. D.Sci. Stoyan Mihov mainly from his work, as well as from the work of his Ph.D. student Assoc. Prof. Dr. Stefan Gerdjikov. I have very good impressions of him as a scientist and a scientific supervisor and I highly appreciate these qualities.

CONCLUSION

In the presented scientific works Assoc. Prof. D.Sci. Stoyan Milkov Mihov has received interesting results in modern fields and has made significant contributions to the theory and its applications. Most of the results have already been used or can be used in similar investigations by other authors. They also have practical application in a number of important areas of informatics and linguistics. Most of the results have been published in respectable issues. I shall add his successful activity as a lecturer and research supervisor, as well as his successful management or participation in important scientific and applied research projects. Although I made some critical remarks, I have every reason to confidently suggest Assoc. Prof. D.Sci. Stoyan Milkov Mihov to take the academic position of "Professor" in the field of higher education 4. Natural Sciences, Mathematics and Informatics, professional field 4.6. Informatics and Computer Science, Specialty Informatics.

September 27, 2021

S[illegible]gnature:
([illegible] BAS)