Институт по информационни и комуникационни технологии-БАН
Вх. №436/27.05.2020г.

REER REVIEW

on the Thesis Submitted for Awarding
the Scientific Degree "Doctor of Sciences"
in the Professional Direction 4.6. Informatics and Computer Science

Author of the thesis: Stoyan Milkov Mihov, Ph.D., Associate Professor in Department of Linguistic Modeling and Knowledge Processing of the Institute of Information and Communication technologies at the Bulgarian Academy of Sciences.

Title of the thesis: „Finite-State Automata, Transducers and Bimachines: Algorithmic Constructions and Implementations".

Member of the Scientific Jury: Doctor of Sciences Vesselin Stoyanov Drensky, Professor at the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, Full Member of the Bulgarian Academy of Sciences.

The thesis is in an area at the meeting point of mathematics and theoretical computer science. An essential part of the thesis studies mathematical problems in the theory of formal languages and automata theory which are realized in a specially created programming language and applied for the solution of practical problems. The thesis is written in English on 226 pages. It contains a short introduction on 3 pages, 8 chapters, an abstract of the contributions, a list of references used in the text and a list of publications of the author in the thesis. The exposition is illustrated with 38 figures. It follows verbatim the first 8 chapters of the book published in 2019 by the author in collaboration with Klaus Schulz. The difference with the book is that the author has omitted the summing up and the exercises in the end of each chapter.

1. **Actuality of the problems studied in the thesis.** 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. The presented thesis is motivated by such kind of problems. The author combines two approaches. From one side finite-state automata and related mathematical objects as finite-state transducers and bimachines are studied from purely mathematical point of view. But then the obtained results are used for creating of working implementations with explicitly realized program products. In particular, especially for the applications the programming language $C(M)$ is created and used to write the program products. In the end of the thesis the author presents explicit applications for important problems in the field. I think that the thesis studies and gives solutions of important and actual problems which have their origin in mathematical linguistics but are also of independent mathematical interest. The author presents also a linguistic interpretation of most of the obtained results. To achieve these results the author uses both already existing methods and

creates a number of new algebraic, combinatorial and computational techniques overcoming serious principal and technical difficulties.

2. **Scientific contributions.** The first chapter "Formal preliminaries" introduces many of the main mathematical objects used in the sequel – alphabets, words and languages. Special attention is paid on the notion of monoid (i.e. a semigroup with 1).

   The original contributions of the author start in the second chapter "Monoidal finite-state automata". In classical theory of finite-state automata and the related theory of regular languages usually one considers automata and languages based on free monoid, when two words are equal if and only if they coincide graphically. A central role in the theory is played by the theorem of Kleene which states that the class of regular languages coincides with the class of languages realized in terms of finite-state automata. The author of the thesis considers finite-state automata and regular languages based on an arbitrary finitely generated monoid. He proves that the obtained finite-state automata and their languages are homomorphic images of the classical ones and establishes the corresponding generalization of the theorem of Kleene. Then the author presents methods for simplification of the automaton structure which gives a simplified verification whether a word belongs to the automaton language. For me personally would be interesting to know how the defining relations and the algorithmic properties of the monoid influent the algorithmic properties of the automaton and its language but this is a question of another study and is not in the spirit of the problems studied in the thesis.

   In the third chapter "Classical finite-state automata and regular languages" the author develops methods for determinization and minimization of classical finite-state automata. He presents a construction which replaces an arbitrary finite-state automaton over a free monoid with an equivalent deterministic one. Using algebraic methods he proves the existence of a minimal deterministic automation for a given language and presents the corresponding construction. Then the author introduces coloured deterministic finite-state automata. These automata are useful for problems when the set of words of the language has to be divided in several groups. A theorem established in the chapter gives the equivalence of coloured deterministic finite-state automata with minimal not coloured ones. It has turned out that deterministic finite-state automata over an arbitrary monoid do not share a series of properties typical for automata over free monoids. To solve the problem the author introduces the weaker class of pseudo-deterministic monoidal finite-state automata and establishes that these automata realize the same languages.

   The fourth chapter "Monoidal multi-tape automata and finite-state transducers" studies multi-tape finite-state automata which are natural generalizations of classical finite-state automata. Since they are based on Cartesian products of several monoids, they are able to handle relations. It has turned out that many of the properties of classical finite-state automata hold also for the multi-tape ones. The chapter studies also the properties of finite-state transducers which are built on the Cartesian product of two automata one of which is free and of classical finite-state transducers when both factors in the Cartesian product are free. The author offers a procedure which solves the problem for functionality of classical finite-state transducers.

   The fifth chapter "Deterministic transducers" deals with transducers which are deterministic on the input tape. Such transducers are commonly used in the texts processing and speech recognition. As a consequence of the obtained results the author characterizes regular functions

on words of bounded in terms of classical subsequential transducers and gives an efficient procedure for deciding the bounded variation property. Also, a procedure is provided for the minimization of subsequential finite-state transducers. In the end of the chapter the author shows that the obtained results hold also in the case when the output monoid is the additive monoid of the nonnegative integers.

The sixth chapter "Bimachines" studies the class of bimachines. These are deterministic finite-state machines which present the class of regular functions on words. There is a standard construction of bimachines starting with functional transducers. The author suggests a new construction which directly transfers the transducer in a bimachine. For many classes of transducers the constructed bimachine has exponentially less states in comparison with the standard construction.

The seventh chapter "The $C(M)$ language" introduces the developed new programming language $C(M)$ which is used for the realization of the algorithms from the previous chapters of the thesis. The new language has the advantage that the realization of the solution of the given problem formally describes the form of the mathematical object which has to be obtained. As a result one concentrates on the mathematical steps on abstract level instead on details on low level. The compiler for the language $C(M)$ is freely available at the home page of the author.
The eighth chapter "$C(M)$ implementation of finite-state devices" contains realizations on $C(M)$ of the automata constructions in the previous chapters. Every section contains an illustrative application: (1) A program which determines the validity of a give date; (2) A program which checks whether a word enters in a dictionary and gives a number of words from the dictionary which are close to the given word (with respect to the Levenshtein distance); (3) Mapping an integer to its phonetization; (4) The usage of bimachines for the realization of arithmetic operations with unbounded integers.

3. **Analysis of the publications on the thesis.** The thesis is based on a monograph published in 2019 by Cambridge University Press which is one of the world leaders among the scientific publishers, 11 papers and one chapter of a book: 7 papers are published in the period 2000 – 2008, 4 are from the period 2011 – 2017 and one is a preprint in arXiv.org from 2017. All papers are published in recognized journals or proceedings of conferences: in Computational Linguistics – 2, in Theoretical Computer Science – 1, International Journal of Document Analysis and Recognition – 1, ACM Transactions on Speech and Language Processing – 1, LNCS – 3, Trends in Linguistics - Studies and Monographs – 1, in the proceedings of other conferences – 2. Three of the publications are with impact factor (two of them, in Computational Linguistics, are in quartile Q1 and one, in Theoretical Computer Science, is in quartile Q3). Besides these three papers, other 4 papers are with SJR factor. As an example of the quality of the papers I shall mention that more than 70 papers were submitted for the conference where paper No. 1 was published. The first 19 papers were published in LNAI series of Springer. The next 19 papers, including this paper, were published in a supplementary proceedings and are available at the home page of CEUR-WS. My opinion is that 12 papers with the quality of these presented for the thesis is a moderate achievement but it is completely compensated with the monograph in a high level publishing house. The author states that, following SCOPUS, the publications related with the thesis have 227 citations. But it would be better to present a list of the citations which would make easier the work of the members of the Scientific Jury and would give an additional possibility of the audience to estimate the scientific achievements of the author. The same holds for the need of more explicit data for the impact factor and the SJR

factor. All publications are with coauthors: the monograph and 7 of the papers are with one coauthor, 4 of the papers are with two coauthors and one is with three coauthors. Since the documentation does not contain statements of the coauthors of the joint papers I accept that all coauthors have the same contribution. The coauthor of the monograph Klaus Schulz declares that Stoyan Mihov has the leading role in the monograph. The coauthors are Klaus Schulz from Germany (coauthor of the monograph and 7 papers), Stefan Gerdjikov from the Faculty of Mathematics and Informatics of the University of Sofia (FMI of Sofia Univ.), a Ph.D. student of the author (4 papers), Petar Mitankin from FMI of Sofia Univ. (2 papers) and Galia Angelova (IICT), Hristo Ganchev (FMI of Sofia Univ.), Jan Daciuk (Poland), Bruce Watson and Richard Watson (South Africa) Denis Maurel (France) and Christoph Ringlsteller (Canada) – coauthors in one paper. The list of papers with impact factor and SJR factor also contains only joint papers. The joint papers are typical for the field because the investigations have a multidisciplinary character. Additionally, the joint publications show the capability to work in a team and I think that this is very important for a scientist. But I would wish to the author to publish also papers without coauthors and I am convinced that he has the potential for this.

4. **Approbation of the results.** Many of the results included in the thesis have been presented at national conferences in mathematical logic in Bulgaria, at a seminar of the University of Munich in Germany and international conferences in Bulgaria, Sweden, France, Spain and the USA.

5. **Recommendations and critical remarks.** I do not have essential critical remarks on the exposition. It would be convenient if the author was presented more information where are published his concrete achievements. For example he should mention which theorem in which paper appeared. Such an information is available in the thesis but I think that it should be more precise. The approach of the author to follow verbatim the first 8 (of total 11) chapters from his monograph with Schulz has its advantages. This means that the text has passed a serious reviewing before publishing the book. Additionally, one of the goals of the book is to serve as an introduction to the field and this allows to use the thesis for attracting Ph.D. and Master students to this area of science, which is a positive feature. As it is mentioned in the introduction of the book, the examples and the programs have "pedagogical" nature focusing mainly on the transparence of the steps and the simplicity of the programs. This is a big advantage for a book but I think that the thesis has to contain more serious applications. The existence of such applications is mentioned in several remarks where the author informs that he has improvements of the algorithms and programs, giving references to papers included in the thesis. I think that the author knows very well the literature in the field. Three of his papers included in the thesis have more than 30 titles in the list of references. For one of the papers this list has more than 40 titles. But the list of references in the thesis contains only 48 titles and includes publications from 1961 till nowadays. I want to mention that the author is a coauthor of all publications in the list of references which appeared after 2010. Since the thesis will be read separately from the papers used in it, I think that the list of references should be longer. Also, it would be useful if the documentation contained a CV of the author as well as the complete list of his publications and for the list of publications with impact factor and SJR factor the values of the factors.

6. **The abstract of the thesis and the abstract of the contributions** are written sufficiently detailed and give clear and adequate information for the contents and the main results of the thesis.

Conclusion. The presented thesis is in an actively developed area of science. It is at very high scientific level and satisfies all requirements to a thesis in the area of informatics and computer science. I recommend to the respectable Scientific Jury to award Associate Professor Ph.D. Stoyan Milkov Mihov with the scientific degree "Doctor of Sciences" in the Professional Direction 4.6. Informatics and Computer Science.

NOT FOR PUBLIC RELEASE

Sofia, May 27, 2020.

Signature:

(V. Drensky)