

REVIEW

on the application of Assoc. Prof. Dr. Kiril Simov to participate in the competition to occupy the academic position of "Professor" in the professional field

4.6. Informatics and computer science, topic Informatics

written by Prof. Dr.Sc. Galia Angelova, ICT-BAS

The competition for this position is announced in the State Gazette no. 49 (21.06.2019) for the needs of the Institute of Information and Communication Technologies (ICT-BAS), Department "Linguistic Modeling and Knowledge Processing". The only candidate is Assoc. Prof. Dr. Kiril Simov. In accordance with the requirements of the *Rules for the specific conditions for acquiring scientific degrees and occupying academic positions in ICT-BAS*, the candidates for "professorships" in area 4, professional field 4.6 Informatics and computer science, must have more than 100 points in group of indicators „B“, more than 260 points in group of indicators „Г“, at least 140 points in group of indicators „Д“ and at least 150 points in group of indicators „E“. Dr. Simov presents a record on the fulfillment of the minimum requirements of the ZRASRB with the following number of points: 120 for the indicators „B“, 324 for the indicators „Г“, 228 for the indicators „Д“, and 648 for the indicators „E“. I accept the calculated points in each indicator as they are given in the application records. Dr Simov received his doctoral degree in 2006, has occupied the position "Associated Professor" for more than five years, and no plagiarism has been discovered in his publications. In this way the formal requirements of the *Rules for the specific conditions for acquiring scientific degrees and occupying academic positions in ICT-BAS* are met and even significantly exceeded, especially in the indicator „E“.

Short CV of the applicant

Dr. Simov graduated in 1986 with a Master's Degree in Computer Science from the Faculty of Mathematics and Informatics of Sofia University. In 2006, after completing a doctorate in the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences, in the Department of Mathematical Linguistics, he defended his Ph.D. dissertation entitled "Logical means for processing of linguistic knowledge in HPSG". His scientific interests were focused not only on the formalisms for processing linguistic knowledge in computational linguistics but also on the systematic development of language resources for Bulgarian language. Dr Simov was employed at the Central Institute for Software Products and Systems, at the Institute of Mathematics and Mechanics of the Bulgarian Academy of Sciences (BAS), at the Coordination Centre for Informatics and Computing Techniques in BAS and its successors Central Laboratory for Parallel Information Processing and Institute of Parallel Information Processing (IPIP) where he gets habilitation in 2007, and finally at the Institute of Information and Communication Technologies (ICT-BAS) where he currently leads the Department "Linguistic Modeling and Knowledge Processing".

General description of the presented materials

A total of 27 publications in English were submitted for participation in the competition, all of them visible in Scopus. Indicators in group „B“ are covered by eight publications while a group of another 19 publications contribute to the indicators in group „Г“. These papers were published

after the year of habilitation (2007) and have not been presented in previous procedures for obtaining an academic degree or occupying an academic position. Fifteen papers submitted for the competition are published in journals or series with Impact Rank (SJR), and eight are indexed by Web of Science. Although the applicant has 45 publications the Scopus since 2007, he apparently has limited himself to submitting to the competition only the 27 papers that present his most significant results in the last 12 years. It should also be noted that the full list of Dr. Simov's publications is included in the submitted Curriculum Vitae - the total number is 180.

The Citation Report submitted by the applicant lists 38 citations of 8 publications indexed by Scopus. The total number of citations of Assoc. Prof. Simov in Scopus is 277 with h-index 10, and in Google Scholar - 1919 citations with h-index 21. Although Google Scholar includes self-citations, the latter large number and international authors of citing papers show the significance and recognition of Dr. Simov's results all over the world.

All submitted publications are co-authored, with the applicant being the first author in six of them. I accept that the results described in these publications have been achieved with the equal participation of all the authors of the respective works. The co-authorship does not diminish the importance of Assoc. Prof. Simov's achievements, but rather emphasizes his position as a scientist with an internationally recognized expertise in computational linguistics, who collaborates with a large number of co-authors from all over Europe in a hot interdisciplinary field. The large number of quotations from various international sources is another evidence of the importance and relevance of the scientific achievements of Assoc. Prof. Kiril Simov.

Research Results and Achievements

The research contributions of Dr. Simov are related to three sub-areas of Computational Linguistics and Artificial Intelligence: *(i)* development of language resources and language technologies for Bulgarian and other languages (mainly English); *(ii)* development of conceptual resources and semantic technologies; and *(iii)* applications of language and semantic resources and technologies. The original results presented in the publications in the three sub-areas are as follows:

- *In the field of development of language resources and language technologies:* Development of annotation schemes and formats for language resources. Design and implementation of information extraction procedures for creation of new language resources. Several resources have been developed, using this approach: a corpus of political speech in Bulgarian, a valence dictionary of Bulgarian language and a system for transcribing DBpedia URIs from other languages into Bulgarian.

A general representation of linguistic knowledge is proposed, which allows several automatic text processing tasks to be performed simultaneously. This representation is a kind of extended dependency tree that encodes information needed for syntactic and semantic analysis as well as for resolving coreference links. Based on this representation, an innovative syntactic analysis system was designed, using the results of several parsers, by "voting" and choosing between local and global analysis. An algorithm has also been designed to construct the representation of sentence semantics. The algorithm works on the dependency tree for the sentence.

In order to create language technologies for the Bulgarian language, Dr Simov addressed in his research various tasks including grammatical annotation, recognition of named entities, parsing, and their integrated application. The test results were the best ones for Bulgarian at the time the results were published.

- *In the field of development of conceptual resources and semantic technologies:* A model of the ontology-text relation has been created through a terminological lexicon and an annotation grammar. The dictionary defines the basic forms of terms, morphological characteristics and grammatical information about their syntactic behaviour in the text. Strategies are proposed for establishing correspondences between lexicon elements and ontologies in the absence of an equivalent lexical element or concept. A methodology for building domain ontologies is proposed. As the initial stage of manual ontological acquisition is slow and laborious, it is suggested that the accumulation starts from a core of concepts extracted automatically from textual sources including standards or terminology dictionaries. After the definition of key terms and phrases, a syntactic analysis is applied, which allows to extract semantic relations between the identified terms. The transition from keywords to the definition of terms and their coding in a formal language (e.g. OWL) is formally defined.

Among the most recent results of Assoc. Prof. Simov are the proposed algorithms for expanding a knowledge graph. These algorithms work on relations derived from concepts organised as a knowledge graph built on the WordNet semantic network. In this graph, the synsets (a set of synonymous words) are nodes, and the relationships between them are arcs. The graphs constructed in this way are used to resolve semantic ambiguity (Word Sense Disambiguation) through random walk algorithms (Random Walk on Graphs), and due to this reason the connectivity of the graph also determines the quality of processing. Several algorithms are introduced to extract new relations or adapt entire sentences to extend a graph constructed using WordNet. It has been shown experimentally that the use of expanded graphs improves the Word Sense Disambiguation result by about 10%.

- *In the field of applications of language and semantic resources and technologies:* An ontology with about 2,500 concepts in the IT domain has been built for the purposes of e-learning and another ontology in the field of home textiles for the purposes of the textiles industry. In the LT4eL project (Language Technology for E-Learning, funded by the European commission in FP6) a functionality has been developed to expand user queries by reasoning on the IT ontology. This expansion has been shown to improve search about three times compared to full-text search for all seven languages considered in the project. In another project, adequate vocabulary and grammar have been created by analysing texts related to Bulgarian iconography; the result was used to annotate the descriptions in a collection of Bulgarian icons. An ontology-based dictionary for the Bulgarian language was created, later expanded to the present Bulgarian WordNet - BTB-WN, which currently contains around 22,000 concepts.

A combination of several freely accessible databases with Linked Open Data has been developed (including DBpedia, Geonames, Freebase and some others) as a knowledge resource that facilitates consistent reasoning. Prof. Simov's contribution is the definition of a unifying ontology that allows access to all data in the integrated database.

Two question-answering systems have been developed, using the language technologies described above. Dr. Simov's contribution is in defining the architecture of the systems and the characteristics for assessing the answer correctness.

Using the ontology-text relation proposed by Dr. Simov, an open access database of approximately 650,000 legal documents was created, linked through EuroVoc and Geonames ontologies. The contribution of Dr. Simov is the creation of the structural ontology for the description of documents and the extraction of RDF facts from the annotated documents.

In addition to the academic contributions, the applicant also presents his plans for future work - development of language technologies using neural networks and deep learning as well as application of language technologies for creating knowledge graphs. These issues are among the hottest topics of Artificial Intelligence today.

Participation in research initiatives: projects, organisation of scientific forums

Assoc. Prof. Kiril Simov, together with Prof. Petya Osenova from the Faculty of Slavic Philology at Sofia University "St. Kl. Ohridski" is extremely consistent in the idea of systematically creating language resources for the Bulgarian language. From 2001 to 2007, in a two-phase project between the University of Tübingen and IPIP-BAS, funded by the Volkswagen Foundation Germany under the "Co-operation with Natural and Engineering Scientists in Central and Eastern Europe" program, he led the development of a Tree Bank of 15,000 syntax trees for the Bulgarian language with annotation in the HPSG formalism. Called BulTreeBank, this resource was a rare artefact for the period 2005-2007 (and an exclusive artefact for a Slavic language), and has attracted the interest of many computational linguists involved in parsing. This bank is accompanied by a morphological dictionary and an annotated text corpus that became the basis of the so-called BLARK (Basic Language Resource Kit). If today the Bulgarian language is equipped with sufficient language resources for basic automatic processing, this is due to the BulTreeBank projects, the dedication of Dr. Simov and the vision IPIP-BAS to take the development of language resources as a priority. From 2008 to 2016, Assoc. Prof. Simov led the Bulgarian team in eight other projects funded by the European Commission's framework programs and has been recognised as one of the few competent computational linguists in the Balkan countries. Dr. Simov is a desirable partner in developments related to multilingual tasks, e.g. machine translation. His projects allowed him to form a group that continues to work to this day on new and complex challenges, such as semantic text analysis, and is the core team for the implementation of the CLADA-BG infrastructure project.

Since 2007 (i.e. after the habilitation for Associate Professor) Kiril Simov, together with Petya Osenova, has organised 10 international scientific forums, among which the prestigious 30th edition of the European Summer School of Logic, Language and Information (ESSLLI 2018) held for two weeks in Sofia in August 2018 with about 400 participants. The extremely high attendance at the ESSLLI 2018 Summer School is another indication of the good organisers' image and the credibility of the international community that Kiril Simov and Petya Osenova will fulfill their commitments to organise a high quality event.

Personal Impressions

I have known Kirill Simov since 1984, when he was one of my first graduates in the Mathematical Linguistics section of the Institute of Mathematics and Mechanics of BAS. His thoroughness, persistence, perseverance and motivation made a big impression. In 1988 he was the first programmer to work seriously on the MorphoAssistant system for morphological analysis and synthesis of Bulgarian words. Today, 30 years later, we see that Assoc. Prof. Simov has retained his interest in creating dictionaries and corpora for the Bulgarian language and continues this activity by developing resources for syntax and semantics.

Over the years, Prof. Simov has developed an understanding and empathy for the work of colleagues in the humanities, especially those related to language studies and cultural heritage. This attitude, infrequent for a computer scientist, helps him a lot in the management of large projects such as CLADA-BG, where most partners come from linguistic units or museums. Prof. Simov's broad professional competence also attracts many candidates for part-time or paid doctoral studies who prefer him as their supervisor (for example, in the field of biomedical ontologies or databases - the PhD students coming from Pensoft or Jens Kohler from the University of Applied Sciences in Mannheim, who defended his doctoral dissertation at IICT in 2018 with Dr. Simov as a research consultant).

Conclusion

Assoc. Prof. Kiril Simov is a leading European researcher in the field of computational linguistics, known as the creator of the CLaRK corpus development environment, the BulTreeBank resource, as coordinator of CLARIN-BG and CLADA-BG and organiser of many scientific forums. The materials presented for the competition prove his profound knowledge of language and semantic technologies, his capacity for project management, his ability to form teams, his perseverance in attracting doctoral students - all of these qualities that are supposed to be inherent for a "Professor" at the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences. **I strongly support the selection of Assoc. Prof. Kiril Simov as a Professor in the Linguistic Modeling and Knowledge Processing Department of IICT-BAS and propose that the members of the Scientific Jury unanimously vote in favor of such a decision.**

16 October 2019

Sofia

Member of the Scientific Jury:

**NOT FOR
PUBLIC RELEASE**

Prof. Dr.Sc. Galia Angelova