SIXTH
INTERNATIONAL
CONFERENCE

COMPUTATIONAL LINGUISTICS IN BULGARIA CLIB 2024

9 – 10 September 2024
Sofia, Bulgaria

PROCEEDINGS

Department of Computational Linguistics
Institute for Bulgarian Language
BULGARIAN ACADEMY OF SCIENCES

The Sixth International Conference *Computational Linguistics in Bulgaria* (CLIB 2024) is organised with the support of the National Science Fund of the Republic of Bulgaria under Grant Agreement No. KP-06-MNF/5 from 23.05.2024.

NATIONAL
SCIENCE
FUND
Ministry of Education and Science

The National Science Fund is not responsible for the content of the papers presented at the Conference or for the content of the promotional and other Conference materials.

CLIB 2024 is organised by:

Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences

PUBLICATION AND CATALOGUING INFORMATION

Title:	Proceedings of the Sixth International Conference *Computational Linguistics in Bulgaria* (CLIB 2024)
ISSN:	2367 5675 (online)
Published and distributed:	Institute for Bulgarian Language Bulgarian Academy of Sciences
Editorial address:	Institute for Bulgarian Language Bulgarian Academy of Sciences 52 Shipchenski Prohod Blvd., Bldg. 17 Sofia 1113, Bulgaria +3592/ 872 23 02

Proceedings of the

Sixth International Conference

Computational Linguistics in Bulgaria

9 – 10 September 2024
Sofia, Bulgaria

PROGRAMME COMMITTEE

Chair:

Svetla Koeva – Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria

Co-chair:

Petya Osenova – Sofia University, Faculty of Slavic Studies; Institute of Information and Communication Technologies, Department of AI and Language Technologies, Bulgarian Academy of Sciences, Bulgaria

Iana Atanassova – University of Burgundy, Centre for Interdisciplinary and Transcultural Research, France

Verginica Barbu Mititelu – Research Institute for Artificial Intelligence, Romanian Academy, Romania

Khalid Choukri – Evaluations and Language Resources Distribution Agency, France

Ivan Derzhanski – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

Tsvetana Dimitrova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

Milena Dobreva – University of Strathclyde, United Kingdom

Radovan Garabík – Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia

Maria Gavrilidou – Institute for Language and Speech Processing, Natural Language Processing and Knowledge Extraction Department, Greece

Voula Giouli – Institute for Language and Speech Processing, ATHENA Research Centre, Greece

Momchil Hardalov – Amazon AWS AI Labs, Spain / Sofia University St. Kliment Ohridski, Bulgaria

Radu Ion – Research Institute for Artificial Intelligence, Romanian Academy, Romania

Ivan Koychev – Sofia University, Faculty of Mathematics and Informatics, Bulgaria

Cvetana Krstev – University of Belgrade, Faculty of Philology, Serbia

Eric Laporte – University of Paris-Est Marne-la-Vallée, France

Natalia Loukachevitch – Research Computing Center of Moscow State University, Russia

Sussi Olsen – University of Copenhagen, Centre for Language Technology, Denmark

Ana Ostroski Anic – Institute of the Croatian Language, Croatia

Vasile Pais – Research Institute for Artificial Intelligence, Romanian Academy, Romania

Bolette Pedersen – University of Copenhagen, Centre for Language Technology, Denmark

Maciej Piasecki – Wrocław University of Technology, Poland

Georg Rehm – German Research Center for Artificial Intelligence, Germany

Ewa Rudnicka – Wrocław University of Technology, Poland

Inguna Skadina – Tilde / Institute of Mathemacs and Computer Science, University of Latvia, Latvia

Milena Slavcheva – Institute of Information and Communication Technologies, Department

ORGANISING COMMITTEE

Table of Contents

PLENARY TALKS

Large Language Models for the Real World: Explorations of Sparse, Cross-lingual Understanding and Instruction-Tuned LLMs

Dr. Veselin Stoyanov (Tome AI, USA)

Large language models (LLMs) have revolutionized NLP and the use of Natural Language in products. Nonetheless, there are challenges to the wide adoption of LLMs. In this talk, I will describe my explorations into addressing some of those challenges. I will cover work on sparse models addressing high computational costs, multilingual LLMs addressing the need to handle many languages, and work on instruction finetuning addressing the alignment between model outputs and human needs.

Ten Years of Universal Dependencies

Prof. Joakim Nivre (Uppsala University and RISE, Sweden)

Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. Since UD was launched almost ten years ago, it has grown into a large community effort involving over 500 researchers around the world, together producing treebanks for 148 languages and enabling new research directions in both NLP and linguistics. In this talk, I will review the history and development of UD and discuss challenges that we need to face when bringing UD into the future.

Written Text Processing and the Adaptive Reading Hypothesis

Prof. Vito Pirrelli (NRC, Institute for Computational Linguistics, Pisa, Italy)

Oral reading requires the fine coordination of eye movements and articulatory movements. The eye provides access to the input stimuli needed for voice articulation to unfold at a relatively constant rate, while control on articulation provides internal feedback to oculomotor control for eye movements to be directed when and where a decoding problem arises.

A factor that makes coordination of the eye and the voice particularly hard to manage is their asynchrony. Eye movements are faster than voice articulation and are much freer to scan a written text forwards and backwards. As a result, given a certain time window, the eye can typically fixate more words than the voice can articulate.

According to most scholars, readers compensate for this functional asynchrony by using their phonological buffer, a working memory stack of limited temporal capacity where fixated words can be maintained temporarily, until they are read out loud. The capacity of the phonological buffer thus puts an upper limit on the distance between the position of the voice and the position of the eye during oral text reading, known as the eye-voice span.

In my talk, I will discuss recent reading evidence showing that the eye-voice span is the "elastic" outcome of an optimally adaptive viewing strategy, interactively modulated by individual reading skills and the lexical and structural features of a text. The voice span not only varies across readers depending on their rate of articulation, but it also varies within each reader, getting larger when a larger structural unit is processed. This suggests that skilled readers can optimally coordinate articulation and fixation times for text processing, adaptively using their phonological memory buffer to process linguistic structures of different size and complexity.