

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.



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 <i>Computational Linguistics in Bulgaria</i> (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
Copyright:	Copyright of each paper stays with the respective authors. The works in the Proceedings are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0).  License details: http://creativecommons.org/licenses/by/4.0 Copyright © 2024

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 Dobрева – University of Strathclyde, United Kingdom

Radovan Garabfk – 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 Mathematics and Computer Science, University of Latvia, Latvia

Milena Slavcheva – Institute of Information and Communication Technologies, Department

of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences, Bulgaria

Kresimir Sojat – University of Zagreb, Faculty of Humanities and Social Sciences, Croatia

Ranka Stankovic – University of Belgrade, Serbia

Ivelina Stoyanova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

Stan Szpakowicz – University of Ottawa, Canada

Hristo Tanev – Joint Research Centre of the European Commission, Italy

Irina Temnikova – Big Data for Smart Society Institute (GATE), Bulgaria

Tinko Tinchev – Sofia University, Faculty of Mathematics and Informatics, Bulgaria

Maria Todorova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

Cristina Vertan – University of Hamburg, Germany

Shuly Wintner – University of Haifa, Department of Computer Science, Israel

Piek Vossen – Free University of Amsterdam, the Netherlands

Katerina Zdravkova – University St Cyril and Methodius in Skopje, North Macedonia

ORGANISING COMMITTEE

Chair:

Svetlozara Leseva – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

Atanas Atanasov – Sofia University, Faculty of Slavic Studies, Bulgaria

Rositsa Dekova – Plovdiv University, Faculty of Philology, Department of English Studies, Bulgaria

Nevena Grigorova – Gate Institute, Sofia University, Bulgaria

Dimitar Hristov – Cleversoft, Bulgaria

Nikolaos-Digenis Karagiannis – Identrics, Bulgaria

Hristina Kukova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

Todor Lazarov – New Bulgarian University, Bulgaria

Viktoria Petrova – A1 Bulgaria, Bulgaria

Valentina Stefanova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

Ekaterina Tarpomanova – Sofia University, Faculty of Slavic Studies, Bulgaria

Table of Contents

PLENARY TALKS	1
Dr. Veselin Stoyanov <i>Large Language Models for the Real World: Explorations of Sparse, Cross-lingual Understanding and Instruction-Tuned LLMs</i>	2
Prof. Joakim Nivre <i>Ten Years of Universal Dependencies</i>	3
Prof. Vito Pirrelli <i>Written Text Processing and the Adaptive Reading Hypothesis</i>	4
MAIN CONFERENCE	5
Radu Ion, Verginica Barbu Mititelu, Vasile Păiș, Elena Irimia, Valentin Badea <i>A Cross-model Study on Learning Romanian Parts of Speech with Transformer Models</i>	6
Ekaterina Goliakova, David Langlois <i>What do BERT Word Embeddings Learn about the French Language?</i>	14
Camille Lavigne, Alex Stasica <i>Whisper-TAD: A General Model for Transcription, Alignment and Diarization of Speech</i>	33
Iglika Nikolova-Stoupak, Gaël Lejeune, Eva Schaeffer-Lacroix <i>Contemporary LLMs and Literary Abridgement: An Analytical Inquiry</i>	39
Milica Ikonić Nešić, Saša Petalinkar, Mihailo Škorić, Ranka Stanković, Biljana Rujević <i>Advancing Sentiment Analysis in Serbian Literature: A Zero and Few-Shot Learning Approach Using the Mistral Model</i>	58
Lyuboslav Karev, Ivan Koychev <i>Generating Phonetic Embeddings for Bulgarian Words with Neural Networks</i>	71
Nelda Kote, Rozana Rushiti, Anila Çepani, Alba Haveriku, Evis Trandafilii, Elinda Kajo Meçe, Elsa Skënderi Rakipllari, Lindita Khanari, Albana Deda <i>Universal Dependencies Treebank for Standard Albanian: A New Approach</i>	80
Verginica Barbu Mititelu, Tudor Voicu <i>Function Multiword Expressions Annotated with Discourse Relations in the Romanian Reference Treebank</i>	90
Atanas Atanasov <i>Dependency Parser for Bulgarian</i>	98
Madalina Chitez, Ana-Maria Bucur, Andreea Dinca, Roxana Rogobete <i>Towards a Romanian Phrasal Academic Lexicon</i>	106
Laura Rituma, Gunta Nešpore-Bērzkalne, Agute Klints, Ilze Lokmane, Madara Stāde, Pēteris Paikens <i>Classifying Multi-Word Expressions in the Latvian Monolingual Electronic Dictionary Tēzauris.lv</i>	113
Laura Occhipinti <i>Complex Word Identification for Italian Language: A Dictionary-based Approach</i>	119
Ivana Brač, Matea Birtić <i>Verbal Multiword Expressions in the Croatian Verb Lexicon</i>	130
Alessandro Lento, Andrea Nadalini, Marcello Ferro, Claudia Marzi, Vito Pirrelli, Tsvetana Dimitrova, Hristina Kukova, Valentina Stefanova, Maria Todorova, Svetla Koeva <i>Assessing Reading Literacy of Bulgarian Pupils with Finger-tracking</i>	140

Denitza Kurshumova <i>Educational Horizons: Mapping the Terrain of Artificial Intelligence Integration in Bulgarian Educational Settings</i>	150
Ekaterina Tarpomanova <i>Evidential Auxiliaries as Non-reliability Markers in Bulgarian Parliamentary Speech</i>	157
Iglika Nikolova-Stoupak, Eva Schaeffer-Lacroix, Gaël Lejeune <i>Extended Context at the Introduction of Complex Vocabulary in Abridged Literary Texts</i>	166
Junya Morita <i>Corpus-based Research into Derivational Morphology: A Comparative Study of Japanese and English Verbalization</i>	178
Ivan Derzhanski, Olena Siruk <i>The Verbal Category of Conditionality in Bulgarian and its Ukrainian Correspondences</i>	187
Natalia Dankova <i>Lexical Richness of French and Quebec Journalistic Texts</i>	196
Maria Khokhlova, Mikhail Koryshev <i>A Corpus of Liturgical Texts in German: Towards Multilevel Text Annotation</i>	201
Valentin Zmiycharov, Todor Tsonkov, Ivan Koychev <i>EurLexSummarization – A New Text Summarization Dataset on EU Legislation in 24 Languages with GPT Evaluation</i>	206
Petya Osenova <i>On a Hurler Resource for Bulgarian</i>	214
Fabio Maion, Tsvetana Dimitrova, Andrej Bojadziev <i>Unified Annotation of the Stages of the Bulgarian Language. First Steps</i>	220
Amal Haddad Haddad, Damith Premasiri <i>ChatGPT: Detection of Spanish Terms Based on False Friends</i>	227
Jordan Kravev <i>Deep Learning Framework for Identifying Future Market Opportunities from Textual User Reviews</i>	241
Ruslana Margova, Bastiaan Bruinsma <i>Look Who’s Talking: The Most Frequently Used Words in the Bulgarian Parliament 1990-2024</i>	249
Sabrina Mennella, Maria Di Maro, Martina Di Bratto <i>Estimating Commonsense Knowledge from a Linguistic Analysis on Information Distribution</i>	257
Georgi Pashev, Silvia Gaftandzhieva <i>Pondera: A Personalized AI-Driven Weight Loss Mobile Companion with Multidimensional Goal Fulfillment Analytics</i>	264
Stanislav Penkov <i>Mitigating Hallucinations in Large Language Models via Semantic Enrichment of Prompts: Insights from BioBERT and Ontological Integration</i>	272
Maria Todorova <i>Commercially Minor Languages and Localization</i>	277
SPECIAL SESSION ON WORDNETS, FRAMENETS AND ONTOLOGIES	285
Ivelina Stoyanova <i>Semantic features in the automatic analysis of verbs of creation in Bulgarian and English</i>	286
Svetlozara Leseva <i>‘Dipdive’ into Motion: Exploring Lexical Resources towards a Comprehensive Semantic and Syntactic Description</i>	296
Ivelina Stoyanova, Hristina Kukova, Maria Todorova, Tsvetana Dimitrova <i>Multilingual Corpus of Illustrative Examples on Activity Predicates</i>	309
Svetla Koeva <i>Large Language Models in Linguistic Research: the Pilot and the Copilot</i>	319
LIST OF AUTHORS	332

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.