

LREC-COLING 2024

**The 5th RaPID Workshop:
Resources and Processing of linguistic,
para-linguistic and extra-linguistic
Data from people with various forms of
cognitive/psychiatric/developmental impairments**

Workshop Proceedings

Editors

Dimitrios Kokkinakis, Kathleen C. Fraser,
Charalambos K. Themistocleous, Kristina Lundholm Fors,
Athanasios Tsanas, Fredrik Öhman

21 May, 2024
Torino, Italia

**Proceedings of the LREC 2024 workshop on:
Resources and Processing of linguistic, para-linguistic and
extra-linguistic Data from people with various forms of
cognitive/psychiatric/developmental impairments
(RaPID-5 2024)**

Copyright ELRA Language Resources Association (ELRA), 2024
These proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-11-1
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

With the support of:

- SpråkbankenText (SBX)
- The Centre for Ageing and Health (AGECAP)
- The Swedish Parkinson Foundation (Parkinsonfonden)
- The Swedish national infrastructure supporting digital and experimental research (HUMINFRA)

SPRÅKBANKENTEXT
A research infrastructure for language data
and a language technology research unit



Parkinsonfonden



Message from the General Chair

Welcome to the LREC2024 Workshop on *"Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments"* (RaPID-5). This volume documents the Proceedings of the RaPID-5 Workshop, held on Tuesday, May 21st, 2024, as part of the 14th edition of the LREC 2024 conference (International Conference on Language Resources and Evaluation); this year joint with the 30th International Conference on Computational Linguistics (COLING) - *LREC-COLING 2024*.

RaPID-5 aims to be an interdisciplinary forum for researchers to share information, findings, methods, models and experience on the collection and processing of data produced by people with various forms of mental, cognitive, neuropsychiatric, or neurodegenerative impairments, such as aphasia, dementia, autism, bipolar disorder, Parkinson's disease, or schizophrenia. Like the previous four editions, the RaPID-5 workshop's focus is on creation, processing, and application of data resources from individuals at various stages of these impairments and with varying degrees of severity. Creation of resources includes e.g. annotation, description, analysis, and interpretation of linguistic, paralinguistic and extra-linguistic data (such as spontaneous spoken language, transcripts, eye tracking measurements, wearable and sensor data, etc). Processing is done to identify, extract, correlate, evaluate and disseminate various linguistic or multimodal phenotypes and measurements, which then can be applied to aid diagnosis, monitor the progression, or predict individuals at risk.

RaPID-5 invited submissions of papers in all of the aforementioned research areas, particularly emphasizing the multidisciplinary aspects of processing such data and the interplay between clinical, nursing, medical sciences, language technology, computational linguistics, natural language processing/artificial intelligence (NLP/AI), and computer science. The workshop serves as a catalyst for discussing several ongoing research questions that drive both current and future research endeavours, by bringing together researchers from diverse communities. The workshop invited papers describing original research, preferably presenting substantial and completed work, while also welcoming contributions such as negative results, interesting application nuggets, software packages / tools / platforms, small works, or works in progress. It stimulated discussions on various ongoing research questions and challenges by uniting researchers from different communities. We extend our gratitude to the members of the Scientific Program Committee (SPC) for their diligent efforts in reviewing and evaluating all submissions. Each submission received between 2 to 4 reviews, aiding authors in revising and improving their papers accordingly.

There were 11 contributions accepted for the workshop.

Keynote speakers of RaPID-5 were:

- **Dr. Alexandra König**, BSc MSc PhD, Institut national de recherche en informatique et en automatique (INRIA); Cobtek (Cognition; Behaviour; Technology) Lab; University Côte d'Azur, France; and,
- **Prof. Maria Liakata**, EPSRC/UKRI Turing Institute AI fellow, Queen Mary University of London, UK

Workshop URL: <https://spraakbanken.gu.se/en/rapid-2024>.

Topics of interest:

The topics of interest for the workshop session included but were not limited to:

- Guidelines, methods and protocols for (remote) data collection and/or annotation (schemas, tools)
- Infrastructure for the domain: building, adapting and sharing of linguistic resources, data sets and tools
- Acquisition and combination of novel data samples; including digital biomarkers, continuous streaming, monitoring and aggregation of measurements; as well as self-reported behavioral and/or physiological and activity data
- Addressing the challenges of representation, including dealing with data sparsity and dimensionality issues, feature combination from different sources and modalities
- Domain adaptation of NLP/AI tools
- Acoustic/phonetic/phonologic, syntactic, semantic, pragmatic and discourse analysis of data; including modeling of perception (e.g. eye-movement measures of reading) and production processes (e.g. recording of the writing process by means of digital pens, keystroke logging etc.); use of gestures accompanying speech and non-linguistic behavior
- Use of wearable, vision, and ambient sensors or their fusion for detection of cognitive disabilities or decline
- (Novel) Modeling and deep / machine learning approaches such as:
 - multimodal learning
 - large pre-trained Transformer language models [LLMs]
 - explainable and interpretable AI models
- ... for early diagnostics, (severity) prediction, monitoring, classification
- Evaluation of the significance of features for screening and diagnostics
- Evaluation of tools, systems, components, metrics, applications and technologies including methodologies making use of NLP/AI; e.g. for predicting clinical scores from (linguistic and/or digital) features
- Digital platforms/technologies for cognitive assessment and brain training
- Evaluation, comparison and critical assessment of resources
- Involvement of medical/clinical professionals and patients
- Ethical, gender bias, legal and safety questions in research with human data in the domain, and how they can be handled
- Deployment, assessment platforms and services as well as innovative mining approaches that can be translated to practical/clinical applications
- Experiences, lessons learned and the future of NLP/AI in the area

Organizing Committee

- Dimitrios Kokkinakis, University of Gothenburg, Sweden (*Workshop chair*)
- Kathleen C. Fraser, National Research Council, Canada
- Charalambos K. Themistocleous, University of Oslo, Norway
- Kristina Lundholm Fors, University of Lund, Sweden
- Athanasios Tsanas, The University of Edinburgh, UK
- Fredrik Öhman, University of Gothenburg and Sahlgrenska University Hospital, Sweden

Scientific Programme Committee (in alphabetic order)

- Malin Antonsson University of Gothenburg, Sweden
- Chiara Barattieri di San Pietro, IUSS University School for Advanced Studies, Italy
- Visar Berisha, Arizona State University, USA
- Sunghye Cho, University of Pennsylvania, USA
- Gaël Dias, Université de Caen Normandie, France
- Jon Andoni Duñabeitia, Universidad Nebrija, Spain and the Arctic University, Norway
- Valantis Fyndanis, University of Technology, Cyprus and University of Oslo, Norway
- Gloria Gagliardi, University of Bologna, Italy
- Leontios Hadjileontiadis, Khalifa University, United Arab Emirates
- Samuel Hollands, University of Sheffield, UK
- Kristy Hollingshead, Florida Institute for Human & Machine Cognition (IHMC), USA
- Christine Howes, University of Gothenburg, Sweden
- Joel MacAuslan, STAR Analytical Services, USA
- Ioanna Markaki, Karolinska Institutet, Sweden
- Martínez-Nicolás Israel, Universidad de Salamanca, Spain
- Ricardo Muñoz Sánchez, University of Gothenburg, Sweden
- Ulla Petti, University of Cambridge, UK
- Emily Prud'hommeaux, Boston College, USA
- Frank Rudzicz, Dalhousie University and Vector Institute, Canada
- Johannes Tröger, DFKI GmbH and ki:elements, Germany
- Spyridoula Varlokosta, National and Kapodistrian University of Athens, Greece
- Åsa Wengelin, University of Gothenburg, Sweden
- Yasunori Yamada, IBM Research, Japan

Table of Contents

<i>Semantic-based NLP techniques discriminate schizophrenia and Wernicke's aphasia based on spontaneous speech</i> Frank Tsiwah, Anas Mayya and Andreas van Cranenburgh	1
<i>Speech Rate and Salient Syllables Position in Spontaneous Speech of Children with Autism Spectrum Disorder</i> Valentina Saccone	9
<i>Cross-Lingual Examination of Language Features and Cognitive Scores From Free Speech</i> Hali Lindsay, Giorgia Albertin, Louisa Schwed, Nicklas Linz and Johannes Tröger	16
<i>Speech and Language Biomarkers of Neurodegenerative Conditions: Developing Cross-Linguistically Valid Tools for Automatic Analysis</i> Iris E. Nowenstein, Marija Stanojevic, Gunnar Örnólfsson, María Kristín Jónsdóttir, Bill Simpson, Jennifer Sorinas Nerin, Bryndís Bergþórsdóttir, Kristín Hannesdóttir, Jekaterina Novikova and Jelena Curcic	26
<i>Automatic Detection of Rhythmic Features in Pathological Speech of MCI and Dementia Patients</i> Marica Belmonte, Gloria Gagliardi, Dimitrios Kokkinakis and Fabio Tamburini	34
<i>Open Brain AI. Automatic Language Assessment</i> Charalambos Themistocleous	45
<i>Exploring the Relationship Between Intrinsic Stigma in Masked Language Models and Training Data Using the Stereotype Content Model</i> Mario Mina, Júlia Falcão and Aitor Gonzalez-Agirre	54
<i>Establishing Control Corpora for Depression Detection in Modern Greek: Methodological Insights</i> Vivian Stamou, George Mikros, George Markopoulos and Spyridoula Varlokosta	68
<i>A Preliminary Evaluation of Semantic Coherence and Cohesion in Aphasic and Non-Aphasic Discourse Across Test and Retest</i> Snigdha Khanna and Brielle C. Stark	77
<i>Harnessing Linguistic Analysis for ADHD Diagnosis Support: A Stylometric Approach to Self-Defining Memories</i> Florian Raphaël Cafiero, Juan Barrios Rudloff and Simon Gabay	87
<i>Crosslinguistic Acoustic Feature-based Dementia Classification Using Advanced Learning Architectures</i> Anna Seo Gyeong Choi, Jin-seo Kim, Seo-hee Kim, Min Seok Back and Sunghye Cho	95

Conference Program

Tuesday, May 21, 2024

09:00–13:00 **Session A**
Chair: Charalambos Themistocleous

09:00–09:10 **Welcome and Introduction**

09:10–09:40 **Invited Speaker 1: Dr. Alexandra König, "Novel Digital Speech Biomarker for Early Detection of Alzheimer's Disease"**

Session Oral:

09:45–10:05 *Semantic-based NLP techniques discriminate schizophrenia and Wernicke's aphasia based on spontaneous speech*
Frank Tsiwah, Anas Mayya and Andreas van Cranenburgh

10:10–10:30 *Speech Rate and Salient Syllables Position in Spontaneous Speech of Children with Autism Spectrum Disorder*
Valentina Saccone

10:30–11:00 *Morning Coffee Break*

Session Oral:

11:00–11:20 *Cross-Lingual Examination of Language Features and Cognitive Scores From Free Speech*
Hali Lindsay, Giorgia Albertin, Louisa Schwed, Nicklas Linz and Johannes Tröger

11:25–11.45 *Speech and Language Biomarkers of Neurodegenerative Conditions: Developing Cross-Linguistically Valid Tools for Automatic Analysis*
Iris E. Nowenstein, Marija Stanojevic, Gunnar örnólfsson, María Kristín Jónsdóttir, Bill Simpson, Jennifer Sorinas Nerin, Bryndís Bergþórsdóttir, Kristín Hannesdóttir, Jekaterina Novikova and Jelena Curcic

11:50–12:10 *Automatic Detection of Rhythmic Features in Pathological Speech of MCI and Dementia Patients*
Marica Belmonte, Gloria Gagliardi, Dimitrios Kokkinakis and Fabio Tamburini

Tuesday, May 21, 2024 (continued)

12:15–12:45 **Questions or comments to the sessions’s presenters**

13:00–14:00 *Lunch Break*

14:00–18:00 **Session B**
Chair: Dimitrios Kokkinakis

14:00–14:40 **Invited Speaker 2: Prof. Maria Liakata, "Longitudinal language processing for dementia"**

Session Oral:

14:45–15:05 *Open Brain AI. Automatic Language Assessment*
Charalambos Themistocleous

15:10–15:30 *Exploring the Relationship Between Intrinsic Stigma in Masked Language Models and Training Data Using the Stereotype Content Model*
Mario Mina, Júlia Falcão and Aitor Gonzalez-Agirre

15:35–15:55 *Establishing Control Corpora for Depression Detection in Modern Greek: Methodological Insights*
Vivian Stamou, George Mikros, George Markopoulos and Spyridoula Varlokosta

16:00–16:30 *Afternoon Coffee Break*

Tuesday, May 21, 2024 (continued)

Session Oral:

- 16:30–16:50 *A Preliminary Evaluation of Semantic Coherence and Cohesion in Aphasic and Non-Aphasic Discourse Across Test and Retest*
Snigdha Khanna and Brielle C. Stark
- 16:55–17:15 *Harnessing Linguistic Analysis for ADHD Diagnosis Support: A Stylometric Approach to Self-Defining Memories*
Florian Raphaël Cafiero, Juan Barrios Rudloff and Simon Gabay

Session Video:

- 17:20–17:40 *Crosslinguistic Acoustic Feature-based Dementia Classification Using Advanced Learning Architectures*
Anna Seo Gyeong Choi, Jin-seo Kim, Seo-hee Kim, Min Seok Back and Sunghye Cho

17:40–17:50 Questions or comments to the sessions's presenters

17:50–18:00 Discussion and Conclusions