

- **17 November 2026**

9h00-9h30 Welcome coffee & introduction

9h30-10h30: [Damiano Spina](#), RMIT University (Australia)

Decolonising Information Retrieval Research

10h30-11h30: [Jussi Karlgren](#), AMD Silo AI (Finland)

How to create culturally relevant instruction data sets and evaluation tests for generative language models

11h30-12h00: Mihaela Goga (to be confirmed)

12h00-14h00: Lunch

14h00-15h00: [Djoerd Hiemstra](#), Radboud University (The Netherlands)

Blinder, it kin wier net! Language Identification on the Web for Low-Resource Languages

15h00-16h00: Tilahun Yeshambel, Addis Ababa University (Ethiopia)

Amharic in the Era of Language Models: Resources, Retrieval, and Open Challenges

16h00-16h15: Poster booster (3 mn each) [call for posters]

16h15-17h15: Coffee, discussions, posters

- **18 November 2026**

9h00-10h00: Rodrigo Agerri, HiTZ Center - Ixa, University of the Basque Country UPV/EHU

10h00-11h00: [Barbara Plank](#), LMU Munich (Germany)

11h00-11h15: Poster boosters

11h15-12h15 - Coffee and posters

12h15-14h00 : Lunch

14h00-15h00 : [Laurent Prévot](#), Aix Marseille Université (France)

What Can Existing Corpora Tell Us About Language Models? Linguistic and Conversational Sensitivity Across Languages

15h00-16h00: [Adam Jatowt](#), University of Innsbruck (Austria)

Creating NLP Benchmarks for Low-resource Languages

16h00-16h15 – Coffee

16h15-17h00 - Round table

- **19 November 2026**

9h00-10h00: [Shu Okabe](#), TUM Heilbronn (Germany)

Two case studies of NLP for endangered languages: parallel sentence mining and interlinear gloss generation.

10h00-11h00: [Senja Pollak](#), Jožef Stefan Institute (Slovenia)

NLP for news analysis in Slovene

11h00-11h15: Coffee break

11h15-12h15: [Eneko Agirre](#), University of the Basque Country UPV/EH (Spain)

Open LLMs for Your Language That Keep Pace with the Frontier: The Case for Basque

12h00-17h00 -Pizza - Hackathon

17h00 – Hackathon results

- **20 November 2026**

Discussion between invitees and hosts

Posters presentations : Call for presentations - send your proposal to cimi26-organizers@irit.fr

Invited speakers

[Damiano Spina](#)

Title: Decolonising Information Retrieval Research

Abstract:

In information retrieval (IR), we focus on characterising systems that aim to satisfy users' information needs. But whose knowledge, perspectives, and values are represented in our test collections, evaluation methodologies, research questions, and approaches? This talk explores what it means to decolonize IR research, drawing on my own ongoing journey as a non-Indigenous researcher working on Aboriginal land in Australia. I will reflect on the assumptions embedded in established IR practices and draw on my experience as Co-Chair of the ACM SIGIR 2026 Low Resource Environments Track to discuss why decolonising practices are particularly important when developing information access technologies for low-resource languages and communities. Rather than proposing definitive answers, the talk aims to open a conversation about how our research practices can create more space for Indigenous perspectives – perspectives that can enrich our understanding of information access not only in low-resource environments, but across the field more broadly.

Bio:

Dr Damiano Spina is a Senior Lecturer in the School of Computing Technologies at RMIT University, Australia, and an Associate Investigator at the Australian Research Council (ARC) Centre of Excellence for Automated Decision-Making and Society (ADM+S). His research sits at the intersection of information retrieval (IR), human–computer interaction (HCI), and responsible AI, focusing on how people interact with and evaluate information access systems. His work explores responsible approaches to information retrieval research and evaluation, including the development of evaluation methods, test collections, and user-centred approaches. He was awarded an ARC DECRA Fellowship to investigate fair and transparent information access in spoken conversational assistants, and has published over 100 peer-reviewed papers in IR and HCI. He is an ACM Distinguished Speaker, Associate Editor of ACM Transactions on Information Systems, and a member of the International Panel on the Information Environment (IPIE). Beyond academia, Damiano is a Mestre in Capoeira, an Afro-Brazilian cultural practice, and a samba composer and musician with Naarm-based band Wombatuque.



<https://www.damianospina.com>

<https://academics.rmit.edu.au/damiano-spina>

Invited speakers

[Jussi Karlgren](#)

Title: How to create culturally relevant instruction data sets and evaluation tests for generative language models

Abstract:

In information retrieval (IR), we focus on characterising systems that aim to satisfy users' information

Bio:

Jussi Karlgren is a mathematically inclined linguist. He has worked on statistical analyses of human language since 1988, mostly at industrial research labs but with frequent visiting positions at academic research institutions. He is a docent (assoc prof) of language technology at University of Helsinki and a Principal AI Scientist at Silo AI, a part of AMD, where he works on generative language models, most notably on evaluating the quality of their output. Besides research he is a board member and technical advisor to a number of companies. Jussi Karlgren is the organiser of the annual ELOQUENT lab for the analysis of generative language model quality.



Invited speakers

[Djoerd Hiemstra](#)

Title: Blinder, it kin wier net! Language Identification on the Web for Low-Resource Languages

Abstract:

Language identification is the task of assigning a language code (e.g. "English") to an unknown piece of text. While language identification may seem like a solved problem on clean text, its performance degrades substantially on noisy web data. We analyze the quality of four open source language identification tools (Resiliparse, FastText, GlotLID and OpenLID) on the the Open Web Index (<https://openwebsearch.eu>) as well as on the public WiLI 2018, and CommonLID 2026 datasets. For the Open Web Index, we manually annotate a random subset, as well as a subset of documents that were automatically identified as "Frisian" by our Resiliparse algorithm. Frisian is one of the official languages of the Netherlands, spoken by approximately 400,000 people; The Frisian Wikipedia contains about 60,000 pages, and as such Frisian may be considered a relatively well-sourced language among minority languages. We show that on the clean WiLI 2018 data, Resiliparse performs well with about 97% accuracy on both high-resource and low-resource languages, showing that Frisian can be classified correctly. On web-extracted data, however, performance drops to 90% on the random subset, 62% on CommonLID, and dramatically to 45% accuracy on the Frisian OWI slice. We propose alternative approaches to language identification that train models on the most frequent features of a language only, thereby removing the need for a lot of training data, and resulting hopefully in better performance on low-resource languages like Frisian. We plan to show preliminary results at the CIMI workshop in Toulouse, maybe with top results for Frisian: Fryslân boppe!

Bio:

Djoerd Hiemstra is professor at Radboud University where he leads the Data Science group. His research interests include topics in information retrieval and natural language processing. He wrote a well-cited PhD thesis "Using Language Models for Information Retrieval". Djoerd has a special interest in technology that may replace Big Tech in the future, such as federated search. As a member of the OpenWebSearch.eu project, Djoerd and his colleagues were responsible for the indexing stage that builds the Open Web Index. Djoerd is active on the European, not-for-profit, federated social media platform Mastodon (<https://idf.social/@djoerd>), and he founded the community-run, diamond open access Information Retrieval Research Journal, IRRJ (<https://irrij.org>).



Invited speakers

[Tilahun Yeshambel](#)

Title: Amharic in the Era of Language Models: Resources, Retrieval, and Open Challenges

Abstract:

Amharic is one of the most widely spoken African languages, yet it remains comparatively under-resourced in natural language processing (NLP) and artificial intelligence. The limited availability of high-quality corpora, annotated datasets, evaluation benchmarks, pretrained models, and language-specific tools presents significant challenges for developing reliable language technologies and AI models for Amharic. These challenges are further amplified by the morphological richness of Amharic and the limited availability of resources for evaluating modern language models. This talk presents recent efforts toward developing resources and retrieval methods for Amharic, with a particular focus on information retrieval and emerging language-model-based approaches. It will first discuss the challenges of Amharic information retrieval and lessons learned from the development of the Amharic Ad-hoc Information Retrieval Test Collection (2AIRC), including the role of morphological representations, root-based retrieval, query expansion, and neural representations. The talk will then introduce ongoing work on AmharicIR+Instr, a resource designed to support neural retrieval and instruction tuning through retrieval-oriented and instruction-following datasets. Beyond these resources, the talk will discuss the broader implications of building AI models for Amharic in the era of large language models. Particular attention will be given to data quality and scale, annotation and evaluation, domain coverage, retrieval-augmented generation, instruction tuning, model evaluation, and the challenges of adapting or developing language models for genuinely low-resource settings. The talk will conclude by identifying open research challenges and opportunities for building more effective, reproducible, and accessible AI resources for Amharic, while highlighting the importance of collaborative resource construction and shared infrastructure for advancing low-resource language technologies.

Bio:

Tilahun Yeshambel is a lecturer at HiLCoE School of Computer Science and Technology, Ethiopia. His research interests focus on information retrieval, natural language processing, low-resource language resource construction and evaluation, machine learning, and neural networks. His work focuses particularly on advancing language technologies for Amharic, including the design and evaluation of Amharic-specific information retrieval models, investigation of morphological and learned text representations, development and evaluation of Amharic language resources and datasets, and training and evaluation of monolingual language models for Amharic. He also develops retrieval benchmarks and annotated datasets for both classical and neural information retrieval, with the aim of contributing reusable resources to the research community and advancing Amharic information retrieval and language modeling.



Invited speakers

[Laurent Prévot](#)

Title: What Can Existing Corpora Tell Us About Language Models? Linguistic and Conversational Sensitivity Across Languages

Bio:

Laurent Prévot is a professor in Language Sciences at Aix Marseille Université. He is also an honorary member of the Institut Universitaire de France (IUF). His specialties are linguistics (in particular semantic and pragmatic questions) and Natural Language Processing. He obtained his PhD (2004) in Computer Science from Toulouse 3 University and worked as post-doc in linguistic and interdisciplinary labs in Italy (Laboratory for Applied Ontology, Trento), France (Syntax and Semantics Research Team CLLE-ERSS, Toulouse 2) and Taiwan (Institute of Linguistics, Academia Sinica). His main research focus are “conversation” and “spontaneous speech” as crucial genres for understanding human language and nature. He takes the success of large language models as a major new fact that linguists need to address. However, he keeps in mind that thousands of languages remain largely excluded from these approaches. He is interested in new models and architectures, and considers that we need a better balance between efforts devoted to data collection and analysis, and those devoted to model training and development.



Contact : Josiane.Mothe@irit.fr