17 November 2026

Horaire	Programme
8h45	Welcome coffee
8h55	Introduction
9h00-10h00	Damiano Spina, RMIT University (Australia) **Decolonising Information Retrieval Research**
10h00-10h30 Short presentations 7 mn + 3mn QA	Silvio Calderaro, Università di Pisa / Università di Napoli "L'Orientale", Italy **How tokenizers used by open-weight language models segment Sardinian texts** Debora Ciminari, University of Bologna, Italy **Research project on Romagnol** Myles Barney, Brigham Young University, USA **Zero-shot efficacy of English-trained Acoustic-to-articulating**
10h30-10h45	Coffee break
10h45-11h45	Jussi Karlgren, AMD Silo AI (Finland) **How to create culturally relevant instruction data sets and evaluation tests for generative language models**
11h45-12h15	Mihaela Goga **Romance-MoLE: Rich Cousins and Their Benefits. An Approach to Low-Resource Language Modelling**
12h15-12h35	Maguelonne Teisseire & Sarah Valentin **Building Annotated «Speech» Corpora for Low-Resource Languages: Methods and Challenges & Towards an AI model for Reunionese Creole. Summary**
12h35-12h45	Farah Benamara **Towards Cultural Aware LLMs in Low-Resources Languages**
12h45-14h00	Lunch Espanade
14h00-15h00	Djoerd Hiemstra, Radboud University (The Netherlands) **Blinder, it kin wier net! Language Identification on the Web for Low-Resource Languages**
15h00-15h30	Tilahun Yeshambel, Addis Ababa University (Ethiopia) **Amharic in the Era of Language Models: Resources, Retrieval, and Open Challenges**
15h30-16h00 Poster booster (3 mn each) QA during the break	Filippos Triantafyllou (180 sec) **Towards cross-domain argument mining** Fang Zhao (180sec) [remote] **Décisions juridiques peu dotées ou l'analyse syntaxique en constituants du moyen néerlandais** Nishchal Prasad, IRIT (180 sec) **Causal Inference with LLMs under Single Observation** Nihel Loubna Ghennani, IRIT (180 sec) **La détection précoce du TSA : un domaine à faibles ressources** Omar Naim (180 sec) **Why Transformers Struggle with Distribution-Independent In-Context Learning** A. Audibert, B. Favre, M. Hjuler et F. Portet (180 sec) **FRAAC: A French Multimodal Large Language Model and Dataset forPictogram-Based Augmentative and Alternative Communication**
16h00-17h00	Coffee, discussions, posters

Contact : Josiane.Mothe@irit.fr

18 November 2026

Horaire	Programme
9h45-10h00	Welcome coffee
10h00-11h00	Barbara Plank, LMU Munich (Germany)
11h00-12h00 Short presentations 7 mn + 3mn QA	Tatiana Anikina, German Research Center for Artificial Intelligence (DFKI) **Synthetic Data Generation for Multilingual Math Reasoning Tasks** Ciré Diallo, CEO, WONHUI, Paris **Script-Weighted tokenizer in Multilingual and Multi-script setting** Sameer Ali, Universitat Oberta de Catalunya **Low-Resource AI for B5G Cybersecurity** Koyel Ghosh, University of Hildesheim **The Multimodal Avenger: Tackling Online Hate Across Languages and Modalities** Ndeye Emilie Mbengue, INRIA Université Côte d'Azur **Classification and formal definition of low resource languages in the semantic web**
12h00-14h00	Lunch - Espanade
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) **The Past Is Not One Domain: Language Evolution, Temporal Adaptation, and Reasoning Across Time**
16h00-16h30	Coffee + poster

19 November 2026

Horaire	Programme
9h00-10h	Shu Okabe, TUM Heilbronn (Germany) **Two case studies of NLP for endangered languages: parallel sentence mining and interlinear gloss generation**
10h-10h15	Coffee break
10h15-11h15	Senja Pollak, Jožef Stefan Institute (Slovenia) **NLP for news analysis in Slovene**
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**
12h15-18h00	Pizza – Hackathon (including a break)
18h00	Hackathon results

Contact : Josiane.Mothe@irit.fr

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

Contact : Josiane.Mothe@irit.fr

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.

Contact : Josiane.Mothe@irit.fr

Mihaela Goga

Title: Romance-MoLE: Rich Cousins and Their Benefits. An Approach to Low-Resource Language Modelling

Abstract:
Integrating low-resource languages into Large Language Models (LLMs) is fundamentally limited by persistent data scarcity and the "Curse of Multilinguality," where dominant source languages overwrite minority language representations and cause structural "translationese". To preserve the endangered Languedocien dialect of Occitan without catastrophic forgetting, we present Romance-MoLE (Mixture of LoRA Experts), a parameter-efficient framework applying targeted cross-lingual transfer from Occitan's phylogenetic "rich cousins": Catalan and French. Using the pre-trained Llama-3.1- Carballo architecture, the pipeline applies Trans-Tokenisation to repair subword fragmentation and utilises curated synthetic data to enforce standardised dialectal norms. Independent Low-Rank Adaptation (LoRA) experts extracted for French and Catalan are fused via TIES-Merging to construct a stable baseline initialisation. Finally, a Hierarchical Mixture of Ranked Adapters routing strategy (HMoRA) applies token-level gating in shallow layers for localised morphological precision and sequence-level pooling in deep layers to eliminate mid-sentence code-switching. Romance-MoLE yields a +14.57 chrF++ improvement in Occitan translation quality over a standard fine-tuning baseline, while mitigating catastrophic forgetting in Catalan and allowing the model to follow French instructions with a significant reduction in structural code-switching.

Bio:
Mihaela Goga is a computational linguist based in Sweden. She holds a Master of Arts in Language Technology from the University of Gothenburg, Sweden. Her research interests focus on multilingual large language models, cross-lingual transfer, the effect of pretraining data on model behaviour, low-resource and minoritised languages, and parameter-efficient adaptation. In particular, her work centers on bringing such languages into LLMs, including cross-lingual transfer from related high-resource languages, tokeniser adaptation, model merging and mixture-of-LoRA-experts architectures, and the evaluation of translation quality, grammaticality and code-switching. In her master's thesis, carried out at AI Sweden, she adapted Llama-3.1 to Languedocien Occitan using its "rich cousins" French and Catalan. As a native speaker of Aromanian, an endangered Eastern Romance language spoken by Vlach communities across the Balkans, she is especially committed to making language technology accessible to small and minoritised language communities, and hopes to extend this work to Aromanian itself in the future.



Contact : Josiane.Mothe@irit.fr

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://irrj.org).

Contact : Josiane.Mothe@irit.fr

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 (2AIRTC), 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.

Contact : Josiane.Mothe@irit.fr

Barbara Plank



Barbara Plank
Professor (Chair) for AI and Computational Linguistics at LMU Munich
b (dot) plank (at) lmu (dot) de

About Me

I am full professor and chair for AI and Computational Linguistics at LMU Munich, Head of the Munich AI and NLP (MaiNLP) lab, and co-director of the Center for Information and Language Processing (CIS). I am also a visiting full professor at the IT University of Copenhagen.

Research Interests

- Learning under sample selection bias (domain adaptation, transfer learning) and annotation bias
- Learning under limited data, continual learning, semi-supervised and weakly-supervised machine learning applied to cross-domain and cross-language natural language processing.
- Multimodal learning, including learning at the language and vision or speech interface and learning from human data
- Fortuitous supervision to enable variety-space aware natural language understanding

Contact : Josiane.Mothe@irit.fr

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 Universit&eacute. 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

Adam Jatowt

Title: The Past Is Not One Domain: Language Evolution, Temporal Adaptation, and Reasoning Across Time

Abstract:
Historical archives contain vast amounts of textual information, yet they remain challenging resources for modern NLP systems. Rather than constituting a single low-resource domain, the past can be viewed as a sequence of evolving, interconnected domains, each characterized by different vocabulary, word meanings, factual knowledge, document characteristics, and availability of annotated data. In this talk, I will introduce a temporal perspective on low-resource NLP, arguing that understanding historical information requires accounting for changes in language, knowledge, and available evidence across time. Drawing on our research, I will discuss how vocabulary and word meanings evolve, how meaningful linguistic periods can be identified, and how time-aware representations can capture these changes. I will then explore the challenges of accessing and understanding historical document collections, including question answering over noisy archival sources, identifying historically significant and currently relevant information, and reconstructing evolving events through timeline summarization. Finally, I will address temporal ambiguity, changes in factual validity, and methods for retrieving and reasoning over time-sensitive information. By connecting these research directions, the talk will demonstrate why treating historical collections as a single static domain is insufficient. I will conclude by highlighting opportunities for temporal domain adaptation, knowledge transfer across historical periods, and data-efficient evaluation, including ongoing community benchmarking efforts on information recency and question answering over web archives.

Bio:



Adam Jatowt is a Professor in the Department of Computer Science and Deputy Head of the Digital Science Center at the University of Innsbruck. He obtained his PhD in Information Science and Technology from the University of Tokyo and worked for 14 years at Kyoto University. His research focuses on information retrieval and natural language processing, with particular emphasis on temporal aspects of information. He is also interested in how one can retrieve and utilize trustworthy information from document archives such as newspaper archives or web archives. Adam is active in organizing shared tasks and evaluation campaigns in IR/NLP including SemEval 2027's task on Reasoning-intensive Retrieval and CLEF 2027' Lab on Evaluating Recency and Change across the Web. He is a recipient of the Friedrich Wilhelm Bessel Research Award from the Alexander von Humboldt Foundation and the International Excellence Fellowship from the Karlsruhe Institute of Technology. Adam is currently an Associate Editor of ACM TOIS, JASIST and IJDL journals.

Contact : Josiane.Mothe@irit.fr

Shu Okabe

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

Abstract:
More than 3,000 languages currently spoken in the world are considered to be endangered. While all of them are underrepresented compared to institutional languages in the current Natural Language Processing (NLP) field, they are not all equally represented. This presentation will highlight two cases within the broad spectrum of NLP support for endangered languages. The talk will first concern endangered languages with some digital presence, for which written texts are available (e.g., Wikipedia). This section will present the task of parallel sentence mining, which aims to find translation pairs between two comparable monolingual corpora. Such resources can then help develop a machine translation system or, more generally, multilingual NLP systems. The second part will focus on languages with little to no online written presence (e.g., languages with no established writing system). In such cases, languages are usually studied by field linguists, who document them by collecting and annotating fieldwork data. Here, NLP can help partly automate the linguistic annotation process. The presentation will cover the task of interlinear gloss generation, which aims to predict morpheme-level annotations (e.g., apple-PL for the word 'pommes' (apples in French)).

Bio:
Shu Okabe has been a Maître de Conférences (~Assistant Professor) in Computer Science at CentraleSupélec in France since November 2026. His work focuses on models and systems for endangered and underrepresented languages. Previously, he was a postdoctoral researcher in Natural Language Processing in Prof. Alexander Fraser's group at the Technical University of Munich (Campus Heilbronn) in Germany. He received his PhD in Computer Science for a thesis on Computational Language Documentation at Université Paris-Saclay in France, supervised by Prof. François Yvon.

Contact : Josiane.Mothe@irit.fr

Senja Pollak

Title: NLP for news analysis in Slovene

Abstract:
The talk will present natural language processing methods for automated news analysis, with a focus on Slovene. We will start with solutions for typical tasks in the media industry: keyword extraction, named entity recognition, sentiment analysis, and automatic news categorisation. We will then look at deeper questions that interest media and discourse researchers: how language and topics in the news change over time, and how computational methods can reveal framing and bias in news coverage.

Bio:



Senja Pollak is a researcher at the Jožef Stefan Institute in Slovenia, where she leads a team working on natural language processing, with a focus on supporting interdisciplinary research in the digital humanities and computational social sciences. She has worked extensively on news analysis: she coordinated the H2020 project EMBEDDIA on cross-lingual news analysis, co-led the ROBACOFI project on comment filtering, led the national project CANDAS on multilingual news discourse analysis, and serves as a work package leader in the LLM4DH project. She is a member of ELLIS Unit Slovenia and is also a work package leader in the ELLIOT project.

Contact : Josiane.Mothe@irit.fr

Eneko Agirre

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

Abstract:
Abstract: Generative AI models are now multilingual, raising new questions about their relative performance across languages and local cultures, specially in communities with less speakers. In this talk I will explore some of those questions and the lessons we learned along the process. Is it possible to instruct LLMs without native or machine translated instructions? Our methodology only requires raw Basque corpora to attain quality close to much larger commercial models, and can be applied to newer open models close to the frontier. In the process, we study catastrophic forgetting, analyse whether fully exploit their multilingual potential when prompted in non-English languages and whether local knowledge can be transferred from a low-resource to a high-resource language. Our evaluation suite was recognised with a best resource paper award at ACL 2024.

Bio:



Eneko Agirre is Full Professor of Informatics and Head of HiTZ Basque Center of Language Technology at the University of the Basque Country, EHU, in San Sebastian, Spain. Visiting researcher or professor at New Mexico State, Melbourne, Southern California, Stanford, New York and Edinburgh Universities. He received the Spanish Informatics Research Award in 2021, is a member of Jakiunde, the Basque academy of sciences, a fellow of the Association of Computational Linguistics (ACL) and a fellow of the European Laboratory for Learning and Intelligent Systems (ELLIS). He was President of ACL's SIGLEX, member of the editorial board of Computational Linguistics, Journal of Artificial Intelligence Research and Action Editor for the Transactions of the ACL. He is co-founder of the Joint Conference on Lexical and Computational Semantics (*SEM). He is a recipient of three Google Research Awards and six best paper awards and nominations, most recent at ACL 2024. Dissertations under his supervision received best PhD awards by EurAI, the Spanish NLP society, the Spanish AI society and the Spanish Informatics Scientific Association.

Contact : Josiane.Mothe@irit.fr

Farah Benamara

Title: Towards Cultural Aware LLMs in Low-Resources Languages

Abstract: In this talk, I will provide an overview of our ongoing projects at Institut de recherche en Informatique de Toulouse (IRIT) on cultural awareness of large language models with a focus on resource creation, data augmentation strategies, dialect prompting and moral reasoning.

Tatiana Anikina

Title: Synthetic Data Generation for Multilingual Math Reasoning Tasks

Abstract: Mathematical reasoning remains challenging for multilingual language models, particularly in low-resource languages such as Sesotho, Zulu, Luganda, Kannada, and Shona, even on grade-school problems. My ongoing research investigates whether small amounts of synthetic supervision can improve mathematical reasoning in low-resource languages while preserving pretrained capabilities. Initial results show that effectiveness of synthetic data depends on the teacher model and the data generation strategy. For example, direct translation is more effective for Qwen-generated data, whereas Gemma can already produce effective training examples through direct generation. I also look into LLM-as-a-Judge evaluations and compare the scores for fluency, coherence, cultural appropriateness, and mathematical validity.

Contact : Josiane.Mothe@irit.fr