

2025 Activity Report

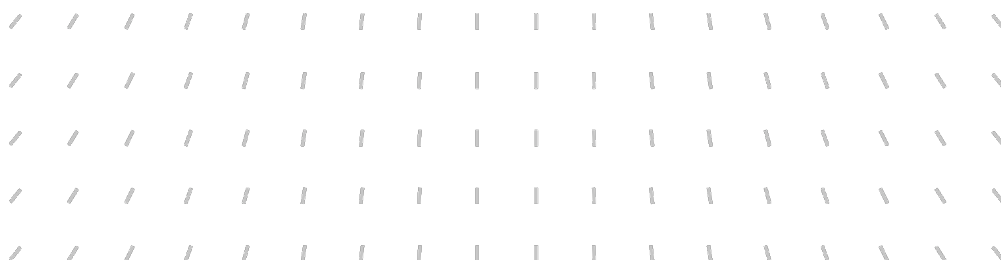
RESEARCH CENTRE: Inria Paris Centre


Project-Team

ALMANACH

Automatic Language Modelling and Analysis &
Computational Humanities





Project-Team ALMANACH

Creation of the Project-Team: 2019 July 01

Each year, Inria research teams publish an Activity Report presenting their work and results over the reporting period. These reports follow a common structure, with some optional sections depending on the specific team. They typically begin by outlining the overall objectives and research programme, including the main research themes, goals, and methodological approaches. They also describe the application domains targeted by the team, highlighting the scientific or societal contexts in which their work is situated. The reports then present the highlights of the year, covering major scientific achievements, software developments, or teaching contributions. When relevant, they include sections on software, platforms, and open data, detailing the tools developed and how they are shared. A substantial part is dedicated to new results, where scientific contributions are described in detail, often with subsections specifying participants and associated keywords. Finally, the Activity Report addresses funding, contracts, partnerships, and collaborations at various levels, from industrial agreements to international cooperations. It also covers dissemination and teaching activities, such as participation in scientific events, outreach, and supervision. The document concludes with a presentation of scientific production, including major publications and those produced during the year.

Keywords

Computer sciences and digital sciences

- A3.1.1. – Modeling, representation
- A3.1.7. – Open data
- A3.1.8. – Big data (production, storage, transfer)
- A3.1.11. – Structured data
- A3.2.2. – Knowledge extraction, cleaning
- A3.2.5. – Ontologies
- A3.3.2. – Data mining
- A3.3.3. – Big data analysis
- A3.4. – Machine learning and statistics
- A3.5. – Social networks
- A3.5.2. – Recommendation systems
- A5. – Interaction, multimedia and robotics
- A5.1. – Human-Computer Interaction
- A5.1.1. – Engineering of interactive systems
- A5.1.2. – Evaluation of interactive systems
- A5.1.7. – Multimodal interfaces
- A5.1.8. – 3D User Interfaces
- A5.1.9. – User and perceptual studies
- A5.6. – Virtual reality, augmented reality
- A5.6.1. – Virtual reality
- A5.6.3. – Avatar simulation and embodiment
- A5.7. – Audio modeling and processing
- A5.7.3. – Speech
- A5.7.4. – Analysis
- A5.7.5. – Synthesis
- A5.8. – Natural language processing
- A9. – Artificial intelligence
- A9.1. – Knowledge
- A9.2. – Machine learning
- A9.2.1. – Supervised learning
- A9.2.2. – Unsupervised learning
- A9.2.3. – Reinforcement learning
- A9.2.4. – Optimization and learning
- A9.2.5. – Bayesian methods
- A9.2.6. – Neural networks
- A9.2.8. – Deep learning
- A9.3. – Signal processing

- A9.4. – Natural language processing
- A9.7. – AI algorithmics
- A9.8. – Reasoning
- A9.10. – Hybrid approaches for AI
- A9.11. – Generative AI
- A9.13. – Agentic AI
- A9.14. – Evaluation of AI models
- A9.15. – Symbolic AI
- A9.16. – Societal impact of AI

Other research topics and application domains

- B1. – Life sciences
 - B1.1. – Biology
 - B1.2. – Neuroscience and cognitive science
 - B1.2.2. – Cognitive science
 - B1.2.3. – Computational neurosciences
 - B2.2.6. – Neurodegenerative diseases
- B2.5. – Handicap and personal assistances
 - B2.5.2. – Cognitive disabilities
- B9.5.1. – Computer science
- B9.5.6. – Data science
- B9.6. – Humanities
 - B9.6.1. – Psychology
 - B9.6.2. – Juridical science
 - B9.6.5. – Sociology
 - B9.6.6. – Archeology, History
 - B9.6.8. – Linguistics
 - B9.6.9. – Political sciences
 - B9.6.10. – Digital humanities
- B9.7. – Knowledge dissemination
 - B9.7.1. – Open access
 - B9.7.2. – Open data
- B9.8. – Reproducibility
- B9.9. – Ethics
- B9.10. – Privacy

Contents

Project-Team ALMANACH	1
1 Team members, visitors, external collaborators	7
2 Overall objectives	10
3 Research program	10
3.1 Research strands	11
4 Application domains	14
4.1 Application domains for ALMAnaCH	14
5 Social and environmental responsibility	14
5.1 Footprint of research activities	14
6 Highlights of the year	16
6.1 Awards	16
7 Latest software developments, platforms, open data	17
7.1 Latest software developments	17
7.1.1 HTR-United	17
7.1.2 CamemBERT-bio	17
7.1.3 RoCS-MT	18
7.1.4 3MT_French Dataset	18
7.1.5 CATMuS Medieval (Model)	18
7.1.6 HTRomance	19
7.1.7 eScriptorium Documentation	19
7.1.8 CATMuS Medieval (Dataset)	19
7.1.9 Counter dataset	19
7.1.10 CamemBERTav2	20
7.1.11 CamemBERTv2	20
7.1.12 LADaS	20
7.1.13 CamemBERT-bio-gliner	20
7.1.14 CoMMuTE	20
7.1.15 mOSCAR	21
7.1.16 Concordancer	21
7.1.17 SWELLS	21
7.1.18 ACReFOSC	22
7.1.19 CoMMA	22
7.1.20 NeWMe	22
7.1.21 SPOT	22
7.1.22 eScriptorium	23
7.1.23 ocDI	23
7.1.24 OcWikiDialects	23
7.1.25 GAPeron	23
7.1.26 ModernCamemBERT	24
7.1.27 Glyphea	24
7.2 New platforms	24
7.3 Open data	25

8	New results	25
8.1	Language Modelling	25
8.1.1	Language Models	25
8.1.2	Multimodal Language Modelling	26
8.1.3	Code Generation	27
8.2	Social, Cultural and Political Computing	27
8.2.1	LLMs' Ability to Understand the Concept of Persuasiveness	27
8.2.2	Evaluating Linguistic Diversity of LLM outputs	28
8.2.3	Political Analysis	28
8.2.4	Cultural Adaptation of LLMs	28
8.2.5	Argumentation in Political Debates	29
8.3	Machine Translation (MT)	29
8.3.1	MT Benchmarking and Shared Tasks	29
8.3.2	MT for Open Science	29
8.3.3	MT for Non-Standard Data	31
8.3.4	Low-resource and Dialectal MT	31
8.3.5	Multimodal MT	33
8.4	Hate Speech and Radicalisation Detection	34
8.4.1	Identifying Common Examples in Spanish Dialects in Hate-Speech Contexts	34
8.4.2	Quantifying Geographically-Informed Sociocultural Biases	34
8.4.3	The CounteR Dataset	35
8.4.4	Inferring Sociological Variation in Algorithmic Text Classification	36
8.4.5	Abusive Language Detection in Online Conversations	36
8.5	Fact-Checking and Disinformation Detection	37
8.6	Dialogue Modelling	37
8.6.1	Repair Modelling	37
8.6.2	Word Meaning Negotiation	38
8.7	Natural Language Processing for Specialised Domains	38
8.7.1	Biomedical and Clinical Domains	39
8.7.2	Patents	40
8.8	Corpus and Tools for Languages of France	40
8.8.1	French OLDI	41
8.8.2	Occitan Dialects	41
8.8.3	Picard	41
8.8.4	French-Based Creoles	41
8.8.5	Language Identification	42
8.8.6	Data Encoding	42
8.8.7	The Parallel Corpus of the Parable of the Prodigal Son	42
8.9	Automatic Text Recognition for Historical Documents	43
8.9.1	Epistemic approach to ATR	43
8.9.2	Improvement and Consolidation of Existing Results	43
8.9.3	New Challenges in ATR	43
8.9.4	Application of ATR at scale	44
8.10	NLP for Historical and Literary Sources	44
8.11	Multimodal Approaches to Human-agent and Human-human Interaction	45
8.11.1	Using Interbrain Synchrony and Rapport Building to Understand Productive Peer Collaboration	45
8.11.2	Son of Sara: Developing a new LLM-based Embodied Conversational Agent	45
8.11.3	Conversational Grounding in Dialogue Systems	46
8.11.4	Exploring Interpersonality: Multimodal Personality Cues in Embodied Conversational Agents	46
9	Bilateral contracts and grants with industry	46
9.1	Active collaborations without a contract	48

10 Partnerships and cooperations	48
10.1 International initiatives	48
10.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program	48
10.1.2 Participation in other International Programs	49
10.2 International research visitors	50
10.2.1 Visits of international scientists	50
10.3 European initiatives	50
10.3.1 Horizon Europe	50
10.4 National initiatives	52
10.4.1 ANR	52
10.4.2 Competitvity Clusters and Thematic Institutes	54
10.4.3 Other National Initiatives	56
10.5 Regional initiatives	63
11 Dissemination	63
11.1 Promoting scientific activities	63
11.1.1 Scientific events: organisation	63
11.1.2 Scientific events: selection	64
11.1.3 Journal	64
11.1.4 Invited talks	65
11.1.5 Scientific expertise	67
11.1.6 Research administration	67
11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach	67
11.2.1 Teaching	67
11.2.2 Supervision	70
11.2.3 Juries	74
11.2.4 Educational and pedagogical outreach	76
11.3 Popularization	77
11.3.1 Productions (articles, videos, podcasts, serious games, ...)	77
11.3.2 Participation in Live events	77
12 Scientific production	77
12.1 Major publications	77
12.2 Publications of the year	79
12.3 Cited publications	86

1 Team members, visitors, external collaborators

Research Scientists

- Benoît Sagot [Team leader, INRIA, Senior Researcher, HDR]
- Rachel Bawden [INRIA, Researcher]
- Justine Cassell [INRIA, Senior Researcher]
- Chloé Clavel [INRIA, Senior Researcher, HDR]
- Thibault Clérice [INRIA, ISFP, from Oct 2025]
- Thibault Clérice [INRIA, Researcher, until Sep 2025]
- Justine Reverdy [INRIA, Starting Research Position, from Sep 2025]
- Carlo Santagiustina [INRIA, ISFP, from Oct 2025]
- Djamé Seddah [INRIA, Researcher, HDR]
- Éric Villemonte De La Clergerie [INRIA, Researcher]

Faculty Member

- Nicolas Rollet [IMT, Associate Professor Delegation]

Post-Doctoral Fellows

- Remy Ben Messaoud [INRIA, Post-Doctoral Fellow, from Sep 2025]
- Elodie Etienne [INRIA, Post-Doctoral Fellow, from Oct 2025]
- Aina Gari Soler [UNIV PSL, Post-Doctoral Fellow, from Oct 2025]
- Aina Gari Soler [INRIA, Post-Doctoral Fellow, until Sep 2025]
- Lucence Ing [INRIA, from Feb 2025]
- Yannis Karmim [INRIA, Post-Doctoral Fellow, from Nov 2025]
- Erinda Morina [INRIA, Post-Doctoral Fellow, from Nov 2025]

PhD Students

- Reem Al Najjar [INRIA]
- Xiangyu An [Afnor, CIFRE, from Jun 2025]
- Wissam Antoun [INRIA]
- Alix Chagué [UNIV MONTREAL]
- Pierre Chambon [META, CIFRE]
- Lucie Chenain [UNIV PARIS - CITE]
- Nicolas Dahan [INRIA]
- Sinem Demirkan [INRIA, from Sep 2025]
- Rasul Jasir Dent [INRIA]

- Romain Froger [Meta, CIFRE, from Jun 2025]
- Matthieu Futeral-Peter [INRIA, until Jul 2025]
- Cecilia Graiff [INRIA]
- Yanzhu Guo [INRIA, until Mar 2025]
- Sofia Imbert De Tremiolles [INRIA, from Nov 2025]
- Francis Kulumba [MINARM]
- Gabrielle Le Bellier [INRIA]
- Simon Meoni [ARKHN]
- Biswesh Mohapatra [INRIA]
- Anh Ha Ngo [INRIA]
- Lydia Nishimwe [INRIA, until Jun 2025]
- Celia Nouri [INRIA]
- Oriane Nédey [INRIA]
- Ziqian Peng [CNRS]
- Arij Riabi [INRIA, until Mar 2025]
- Hugo Scheithauer [INRIA]
- Oussama Silem [INRIA, from Sep 2025]
- Rian Touchent [INRIA]
- Lorraine Vanel [TELECOM PARIS, CIFRE, until Jul 2025]
- Yi Yu [INRIA]
- Armel Zebaze Dongmo [INRIA]
- You Zuo [Questel, CIFRE]

Technical Staff

- Hassen Aguilu [INRIA, Engineer]
- Nicolas Angleraud [INRIA, Engineer, from Mar 2025]
- Barokshana Baskaran [INRIA, Engineer, from Oct 2025]
- Barokshana Baskaran [INRIA, Engineer, from Sep 2025 until Sep 2025]
- Khaled Benaida [INRIA, Engineer, from Oct 2025]
- Sarah Benière [INRIA, Engineer]
- Cindy Evellyn De Araujo Silva [INRIA, Engineer, until Nov 2025]
- Sinem Demirkan [INRIA, Engineer, until Aug 2025]
- Giovanni Duca [INRIA, Engineer, from Sep 2025]
- Nathan Godey [INRIA, Engineer, from Feb 2025 until Apr 2025]

- Zehra Melce Hüsünbeyi [INRIA, Engineer, from May 2025 until Oct 2025]
- Sofia Imbert De Tremiolles [INRIA, Engineer, from Oct 2025 until Oct 2025]
- Juliette Janès [INRIA, Engineer]
- Antonia Karamolegkou [INRIA, Engineer, from Nov 2025]
- Hasan Onur Keles [INRIA, Engineer, until Jun 2025]
- Benjamin Kiessling [INRIA, Engineer, from Jul 2025]
- Maxence Lasbordes [INRIA, Engineer, from Nov 2025]
- Théo Lasnier [INRIA, Engineer, from Oct 2025]
- Marius Le Chapelier [INRIA, Engineer]
- Malik Marmonier [INRIA, Engineer]
- Virginie Mouilleron [INRIA, Engineer]
- Marie Nsingi Kinkela [INRIA, Engineer, from Jun 2025]
- Oussama Silem [INRIA, Engineer, until Aug 2025]
- Panagiotis Tsolakis [INRIA, Engineer]

Interns and Apprentices

- Gabrielle Alimi [INRIA, Intern, until Jun 2025]
- Kshitij Ambilduke [INRIA, Intern, from May 2025 until Aug 2025]
- Barokshana Baskaran [INRIA, Intern, until Jun 2025]
- Clara Coridon [INRIA, Intern, from Mar 2025 until Sep 2025]
- Giovanni Duca [INRIA, Intern, from Mar 2025 until Jul 2025]
- Alix Girard [UNIV PARIS - CITE, Intern, from Nov 2025]
- Sofia Imbert De Tremiolles [INRIA, Intern, from May 2025 until Sep 2025]
- Théo Lasnier [INRIA, Intern, from Apr 2025 until Sep 2025]
- Yassine Machta [INRIA, Intern, from May 2025 until Nov 2025]
- Zofia Milczarek [INRIA, Intern, until Jul 2025]
- Mayank Palan [INRIA, Intern, from May 2025 until Aug 2025]
- Adriano Rivierez [INRIA, Intern, from May 2025 until Aug 2025]
- Imani Symone Stone-Mooring [INRIA, Intern, from Mar 2025 until Jun 2025]

Administrative Assistants

- Sophie Etling [INRIA]
- Martial Le Henaff [INRIA]

Visiting Scientists

- Francesco Benedetti [UNIV PISE, from Sep 2025]
- Marine Carpuat [UNIV MARYLAND, until Jun 2025]

2 Overall objectives

The ALMAnaCH project-team (Automatic Language Modelling and Analysis & Computational Humanities) is a pluridisciplinary team whose research activities are centred around **natural language processing** (hereafter NLP) and **digital humanities** (hereafter DH), and also include **computational linguistics** and **digital social sciences**. Our expertise lies at the crossroads between computer science, machine learning and deep learning, linguistics, and the humanities. ALMAnaCH is the successor at Inria of the project-team ALPAGE (2007-2016), a joint team between Inria and the linguistic department of Université Paris VII (now part of Université Paris Cité), itself a follow-up, for Inria, of the project-team ATOLL.

The evolution of our research field and the arrival of new permanent members have required us to rethink the way we structure our research programme. While preparing ALMAnaCH's 2024 evaluation report, we defined five research axes that bring together our main scientific goals. All five axes are underpinned by the fundamental challenge posed by language variation, in particular through the data-efficiency, the robustness and the adaptability of our models, but also through the need to develop resources for non-standard and low-resource language varieties. Although contemporary, standard French and English will continue to play a central role in our research, we will therefore give a particular importance to non-contemporary and non-standard French as well as to low-resource languages, with a specific focus on languages of France other than French.

3 Research program

As a field, NLP dates back to at least the early 1950s and is one of the main sub-fields of **Artificial Intelligence**. Two decades after the first revolution of the NLP field, when machine learning took over from rule-based approaches in most scenarios, NLP underwent its second revolution in the mid and late 2010's, when the rise of continuous representations [112] enabled deep learning techniques to become dominant. This deep learning era has itself undergone several transformations, in particular with the arrival of masked language models (MLMs) such as BERT [102], which could be fine-tuned to specific tasks, resulting in an impressive performance boost compared to previous methods. Later on, the arrival of large-scale generative models (or large language models, hereafter LLMs) and years after, the rise of conversational models such as ChatGPT revolutionised the perception of what could be a natural interaction with a language agent.

For the first time, the general public can access a range of natural language tools (question answering, machine translation (MT), automatic summarisation, information extraction, even text understanding to some extent) with ease through the simplest unified interface possible, language. Although major questions regarding LLMs and conversational models quickly emerged, in particular concerning their performance for languages other than English—especially minority languages and dialects—and about the cultural biases they might convey, the fact remains that these conversational language models excel at many tasks that used to simply be research topics, at least for edited texts in English and other high-resource languages. This evolution has blurred the lines between research and engineering, without providing a solution to the challenges related to language diversity, language variation and low-resource scenarios in NLP research.

Since ALMAnaCH's creation, our research programme in NLP and DH has been underpinned by these important scientific challenges; as highlighted in our 2019 team creation proposal, our goal has always been **to model and to cope with both language variation¹ and language diversity**. Beyond socio-linguistic factors such as age, gender, origin and education level, language variation arises due to multiple factors such as domain and genre (news wires, scientific literature, oral transcripts, etc.), space and time (geographical

¹Language variation, or language variability, is a term from socio-linguistics, which, as stated by [116] refers to regional, social or contextual differences in the ways that a particular language is used. For example, variations in user-generated content can be characterised through their prevalent idiosyncrasies when compared to canonical texts [106, 121], both across languages and within languages.

variation and evolution over time, cultural differences), which often result in low-resource scenarios. 30 years after the advent of data-driven approaches, which typically rely on supervised and semi-supervised methods requiring large amounts of annotated data, we still often rely on annotated and domain-specific training data to reach the best performance. This is especially the case in specialised domains and low-resource scenarios characterised by a high level of variation, despite the increasing performance levels reached by LLMs, whose training is mostly based on non-annotated data.

Low-resource scenarios also occur as a result of language diversity, for instance when working on multilingual tasks such as MT. Cross-lingual transfer techniques can help dealing with such scenarios, but they cannot fully solve the issue. Language variation and language diversity are not independent challenges and share many characteristics. Dialectal and diachronic diversity form a continuum with language diversity amongst closely-related languages.

With these challenges in mind, ALMAnaCH's central objectives are:

- to develop state-of-the-art NLP techniques, tools and resources to be used by academics and industry;
- to apply our domain of expertise to digital and computational humanities as well as to computational linguistics, both as application domains for NLP and as sources of new NLP-related scientific questions;
- to be dedicated to having an impact on the industry and more generally on society, via collaborations with companies and other institutions (startup creation, industrial contracts, expertise, dissemination, etc.).

3.1 Research strands

Research axis 1: Language modelling

Over the last few years, training a state-of-the-art language model (LM) has gone from a research problem accessible to well-resourced academic labs to a huge engineering effort that only a handful of companies worldwide can carry out. We therefore mostly train LMs in order to investigate scientific questions on how they work and how we can **improve LMs** (new architectures, new units, new loss functions). Having access to all components of these models (dataset, code, parameters) will make it possible to carry out novel experiments on language models. Together with novel approaches such as the study of the formal capabilities of language models, especially large language models (LLMs),² and the use of tailored augmented languages, we plan to develop novel model architectures that will be **more data- and compute-efficient**, and that will have access to **more contextual information**, provided within the left context (“in-context learning”) and externally using tools and agents (see also axis 5).

Another challenge that we will continue to explore, because of its huge importance from a scientific as well as a societal point of view, is the **evaluation of LMs**. Despite a number of large-scale initiatives, designing accurate evaluation protocols and datasets for LLMs, avoiding issues such as data contamination, is still an open problem. With respect to our research axis 3, part of this work will be dedicated to the **development of alignment datasets**³ specifically targeting French and its socio-cultural context. Another important aspect is the evaluation of a model's **trustworthiness**, in particular taking into account the possibility that it was altered, or its training corpus manipulated, including with malevolent intentions (“weaponisation” of LMs). We will develop methods to ensure that our pre-training datasets have not been contaminated with spam, low-quality and adversarial content, possibly LLM-generated (hence the need for LLM-generated content detection models, a task we are already working on).

Research axis 2: Machine translation

Text transformation tasks have become one of our key research directions, our main focus being **MT**. We will focus on three main scenarios. The first one is **MT for scientific documents**, which requires the development of document-level models (e.g. to guarantee document consistency and coherence, including with document-specific elements such as headers, captions and tables). It will also require the development of

²LLMs can be defined as language models with more than 1B parameters.

³Such datasets make it possible to turn a so-called base language model, trained to predict the next token within a text, into a conversational agent trained to interact in a way that is both useful and respectful of certain values.

an interpretable document-level evaluation metric adapted to scientific texts, which will evaluate the quality of term translation, abbreviation handling and co-references. We will also tackle the lack of document-level training data by collecting post-edited versions of machine-translated abstracts, but also by developing data augmentation strategies for longer documents. We plan to integrate our MT models in the **HAL** publishing interface to encourage the submission of abstracts in other languages.

The second scenario is **robust MT**, i.e. MT for non-standard text. We will focus on two types of language variation, **dialectal variation** (translation from or to a dialectal continuum or closely related language varieties) and sociolectal variation as found in **user-generated content (UGC)**. We will develop models that are structurally robust to these variation types (and others), and models that can be controlled, for instance to translate a UGC input into a UGC output with similar characteristics or to translate a text into a specific dialectal variety within a dialectal continuum.

The third scenario we will focus on is **low-resource MT**. We will adapt models to unseen or less represented languages in our training data by **exploiting linguistic analysis** and alternative resources, such as lexicons and grammars. Additionally, we will explore the connections to robustness by identifying **similarities and analogies** between texts from non-standard or rare language varieties and those from well-resourced languages. This approach will facilitate adaptation and extend beyond bilingual lexicon induction to include morphology and grammar.

Transversally to these three scenarios, we will explore the **linguistics of NLP models** to determine how linguistic information can enhance performance. This involves identifying useful contextual examples and understanding how abstract linguistic properties from previous examples can be used to improve predictions on new, complex cases, such as through in-context learning and memory networks. Additionally, we aim to go towards the use of LLMs to **generate educational materials for language learning**, focusing on creating similar example sentences for languages with limited resources. Our research will also concentrate on generating **high-quality and challenging evaluation data**. This will involve developing automated methods to detect such data from existing sources, minimising bias towards certain examples, and improving the detection of context-dependent sentences using more advanced techniques than are currently available.

Research axis 3: Conversational agents

LLMs enable conversational agents to engage users on a range of topics for extended periods. However, human conversation includes unique features often missing in LLMs, such as shorter utterances for listener feedback, complex turn-taking, and multiple concurrent goals such as social bonding and language alignment. We aim to better understand and model human interactions, creating agents that react more naturally and enhance human performance in collaboration and interaction with machines.

Our first research focus is user-oriented, under the prism of the **acceptability**⁴ of current dialogue technologies, mostly conversational neural LMs. Users of a conversational model might not share the same cultural background, values and beliefs, whereas conversational models mostly encode a single cultural viewpoint, creating a **cultural gap**.⁵ Making conversational agents **culturally coherent yet adaptable to a specific user**, as well as more **reliable**, will involve working on model architectures (e.g. disentangling information causing biases, using controllable generation mechanisms, hybridising them with knowledge graphs) and **integrating the humanities and social sciences** at the core of our research.⁶ Another challenge related to the acceptability of conversational models concerns human-machine interaction itself. We aim to understand and model the **lexico-semantic alignment** phenomenon between humans and the way they use **“repair” mechanisms** in dialogue, and to enhance conversational models with such abilities.

Our second research focus is more task-oriented. We will observe conversation among people in a range of ethologically valid contexts, discover and study the essential **characteristics of conversation** (e.g. sentence length, complex turn-taking, use of hedges), model those characteristics, and look at whether particular characteristics predict success on particular collaborative tasks (e.g. human peer tutoring), that allow us to have concrete metrics for the impact of these conversational devices. We will then implement

⁴Acceptability refers to how much users accept to interact with the model and is related to models not displaying undesirable behaviours, opinions and errors.

⁵This is something we have already been studying in the context of the development of radicalisation detection models, especially in the process of creating human-annotated corpora.

⁶The recently started “Inria Action Exploratoire” SaLM, joint with SciencesPo (the leading French higher education institution for sociological and political sciences), led by Djamé Seddah, is dedicated to these topics.

them in **embodied conversational agents**, and carry out experiments to determine whether the presence of these characteristics improve success rates. More recently we have begun collecting more **neurobiological evidence** of the nature of conversation between people, and the impact of those conversational devices on performance. This is carried out through hyperscanning experiments, where we use fNIRS to simultaneously scan the brain waves of two participants in a conversation and extract moments of **inter-brain synchrony**, often thought to be evidence for shared mental models, and the conversational devices they co-occur with.

Research axis 4: Corpus development, computational linguistics and computational humanities

To support the development of NLP models, language documentation and research in digital humanities, we will keep working on the **development of large-scale freely available corpora**. Although web-crawled data is sometimes necessary, we will shift our primary focus from web-based data (e.g. OSCAR) to higher-quality, legally safer data. Firstly, we will build and distribute a **large diachronic corpus of French**, relying on existing data (e.g. the HAL and Persée collections of scientific documents) and on data that we will digitise thanks to our improved OCR/HTR expertise (see below), as well as on an improved processing pipeline, especially improved language detection. We will focus on several low-resource languages, in particular **languages of France other than French** such as Occitan, Alsatian, Picard and French-based Creoles, in the context of the COLaF Inria DEFI. We also aim to support the production of open corpora for **old and ancient languages** to ensure the transition of these languages from closed repositories (such as Thesaurus Linguae Graecae or Library of Latin Texts A/B) to open repositories. For these two sets of low-resource languages we will work on developing **NLP tools and resources** including morphosyntactic annotators, LMs and MT models.

In digital humanities, our long-term goal remains the creation of a complete **pipeline** able to go from layout segmentation and OCR/HTR to publication of the structured digitised corpus in a TEI format, with a focus on **historical documents**. We aim to improve the state of the art in both **layout interpretation** and **text recognition**, based on the eScriptorium platform. For text recognition, we will focus on Latin scripts from the Middle Ages to the contemporary period, but we will also work with other scripts, in particular in the context of the Inria “Action Exploratoire” BackInTime, whose focus is on 17th century **encrypted manuscripts** using custom symbols.

Finally, we will resume our work on computational linguistics, focusing on **computational morphology** and **modelling of language diachrony** (including etymology). These research directions will likely require the development of new **language resources** (e.g. morphological and etymological lexica) or the improvement of existing ones, in relation with our work on corpus development and OCR (e.g. structured lexical information extraction from OCRised dictionaries).

Research axis 5: NLP for specialised domains

We will keep working on specific challenges related to social media text (user-generated content), medical documents, legal texts, administrative documents, patents, employee surveys,⁷ as well as more niche yet scientifically interesting or societally impactful domains (e.g. oenology). A recurring research direction in this regard is the **adaptation of LMs** to specialised domains. Such domains pose various challenges, including domain-specific language variation, in particular terminological and stylistic specificities (e.g. legal style in patents), and access to knowledge bases (e.g. medical ontologies). **Continual pre-training** proved to be successful with CamemBERT-Bio [127, 128], and we will investigate how we can improve the performance of such adaptations, such as using LLM-generated synthetic data and **cross-lingual transfer**.

When we work on **information extraction** and **text generation** systems, their computational efficiency can be a crucial challenge in certain contexts. An example is the **medical domain**, because of the sensitivity, criticality, and confidentiality of patient data required by **small LMs** (SLMs) that can be run in relatively constrained local environments to avoid data leaks. Given the complexity of medical data, one approach is to design **augmented models** that can interact with knowledge databases, combining external calls and extensions of **chain-of-thought** (CoT) processes to handle complex knowledge-based reasoning. The design of such augmented SLMs is still an active area of research and should involve transfer from LLMs (via

⁷This is the activity domain of opensquare, an ALMANACH spin-off co-created in 2016 by Benoît Sagot.

distillation, teacher-student methods, etc.), adaptation to **domain-specific knowledge bases** and **instruction tuning** via reinforcement learning.

4 Application domains

4.1 Application domains for ALMAnaCH

ALMAnaCH's research areas cover Natural Language Processing, a major sub-domain of Artificial Intelligence, as well as Digital Humanities and Computational Social Sciences. Application domains are therefore numerous, as witnessed by ALMAnaCH's multiple academic and industrial collaborations, for which see the relevant sections. Examples of application domains for our research activities include:

- Information extraction, information retrieval, text mining (e.g. opinion surveys)
- Text generation, text simplification, automatic summarisation
- Spelling correction (writing aid, post-OCR, normalisation of noisy/non-canonical texts)
- Machine translation
- Chatbots, conversational agents, question answering systems
- Medical applications (analysis of medical documents, early diagnosis, language-based medical monitoring, etc.)
- Applications in the legal domain
- Applications in linguistics (modelling languages and their evolution, sociolinguistic studies, etc.)
- Digital humanities (exploitation of text documents, for instance in historical research)
- Computational social sciences (e.g. computational analysis of political discourse, social media content analysis, radicalisation detection, etc.)
- Neuroscience

5 Social and environmental responsibility

5.1 Footprint of research activities

Given recent interest into the energy consumption and carbon emission of machine learning models, and specifically of those of language models [122, 98], we have decided to report the power consumption and carbon footprint of all our experiments conducted on the Jean Zay⁸ and Adastral⁹ supercomputers in 2025. For this report, we follow the approach of [124]. While the ALMAnaCH team uses other computing clusters and infrastructures such as CLEPS¹⁰ and NEF,¹¹ these infrastructures are not optimised for large jobs based on multi-node computation, and we therefore consider the power consumption and CO₂ emissions of the experiments conducted in these clusters limited compared to those of Jean Zay and Adastral. Moreover our estimates suppose peak power consumption at all times, which is the worst case scenario and which was clearly not the case at all times for all of our experiments. This could therefore somewhat compensate for the non-reported consumption on both NEF and CLEPS.

⁸Jean Zay documentation

⁹Adastral documentation

¹⁰CLEPS documentations

¹¹NEF documentation

Project ID	Type	Power draw	GPUs	GPU hours	Real hours	Power consumption (kWh)	CO2 emissions
AD011016786	H100	2800	4	1,216	304	1,021	
SS021016138	H100	2800	4	583,985	145,996	490,546	
GC011015610	H100	2800	4	30,949	7,737	25,996	
AD011013900R2	A100	3700	8	4,572	571	2,535	
AD011015138R1	AMD MI300A	3125	4	30,662	7,665	28,743	
AD011015138R1	AMD MI250x	2500	4	10,541	2,635	7,905	
A0161015138	AMD MI300A	3125	4	7,928	1,982	7,432	
A0161015138	AMD MI250x	2500	4	60,236	15,059	45,177	
A0191015138	AMD MI300A	3125	4	4,405	1,101	4,128	
AD011014911	AMD MI250x	2500	4	50,633	12,658	37,974	
AD011015117R2	H100	2800	4	11,199	2,799	9,404	
AD011015117R2	V100	1520	4	4,582	1,145	2,088	
AD011016421	H100	2800	4	299	74	248	
AD011016333	A100	3700	8	2,126	265	1,176	
AD011016333	H100	2800	4	4,174	1,043	3,504	
AD011012254R4	A100	3700	8	5,207	650	2,886	
AD011012254R4	H100	2800	4	23,603	5,900	19,824	
AD011012254R4	V100	1520	4	5,594	1,398	2,549	
AD011015933	A100	3700	8	6,053	756	3,356	
AD011015933	H100	2800	4	17,557	4,389	14,747	
AD011015933	V100	1520	4	2,625	656	1,196	
AD011013674R2	A100	3700	8	10,340	1,292	5,736	
AD011013674R2	H100	2800	4	20,809	5,202	17,478	
AD011013674R2	V100	1520	4	19,732	4,933	8,997	
Total	A100	3700	8	28,298	3,537	15,704	
Total	H100	2800	4	693,791	173,447	582,781	
Total	V100	1520	4	32,533	8,133	14,834	
Total	AMD MI300A	3125	4	42,995	10,748	40,305	
Total	AMD MI250x	2500	4	121,410	30,352	91,056	
Grand Total				919,027		744,680	

Table 1: Project ID, GPU times in hours, real node time in hours, mean power consumption including power usage effectiveness (PUE), and CO₂ emissions for each Jean Zay and Adastra project associated with the team.

Node infrastructure: We have access to three types of GPU node on Jean Zay:¹²

- Nodes comprising 4 Nvidia Tesla V100 SXM2 32GB GPUs, 192GB of RAM, and two Intel Cascade Lake 6248 processors. One Nvidia Tesla V100 card is rated at around 300W,¹³ while the Intel Cascade Lake processor is rated at 150W.¹⁴ For the DRAM we can use the work of [101] to estimate the total power draw of 192GB of RAM at approximately 20W. The total power draw of one Jean Zay node at peak use therefore adds up to around 1520W.
- Nodes comprising 8 Nvidia A100 SXM4 80GB GPUs, 512GB of RAM, and two AMD Milan EPYC 7543 processors. One Nvidia A100 card is rated at around 400W,¹⁵ while the AMD Milan processor is rated at 225W.¹⁶ Following [101], we estimate the total power draw of 512GB of RAM at approximately 50W. The total power draw of one A100 node at peak use therefore adds up to around 3700W.
- Nodes comprising 4 Nvidia H100 SXM5 80GB GPUs, 512GB of RAM and two Intel Xeon Platinum 8468 processors. One Nvidia H100 card is rated around 700W¹⁷, while the Intel Xeon Platinum 8468

¹²Jean Zay architecture description

¹³Nvidia Tesla V100 specification

¹⁴Intel Xeon Gold 6248 specification

¹⁵Nvidia Tesla A100 specification

¹⁶AMD Milan EPYC 7543 specification

¹⁷Nvidia Hopper H100 specification estimate

processor is rated at 300W.¹⁸ The total power draw of one H100 node at peak use therefore adds up to around 3450W.

We have recently gained access to the Adastra supercomputer at CINES,¹⁹ which has

1. nodes comprising 4 AMD MI250Xs and 1 AMD EPYC Trento processor. Adastra uses the same hardware as the LUMI cluster,²⁰ for which the total power draw of one node at peak use has been reported as being approximately 2500W.²¹
2. nodes comprising 4 AMD MI300As (APUs including integrated AMD Zen 4 x86 CPU cores), for which the total power draw of one at node at peak has been reported as being approximately 3,125W²²

With this information, we use the formula proposed by [124] and compute the total power required for each setting:

$$p_t = \frac{1.20t(cp_c + p_r + gp_g)}{1000} \quad (1)$$

Where c and g are the number of CPUs and GPUs respectively, p_c is the average power draw (in W) from all CPU sockets, p_r the average power draw from all DRAM sockets, and p_g the average power draw of a single GPU. We estimate the total power consumption by adding GPU, CPU and DRAM consumption, and then multiplying by the *Power Usage Effectiveness* (PUE), which accounts for the additional energy required to support the compute infrastructure. We use a PUE coefficient of 1.20, which is the value reported by IDRIS for the Jean Zay supercomputer. For the real time t we have to divide the reported time for each Jean Zay project by 4 for V100 and H100 nodes and 8 for A100 nodes, as Jean Zay reports the computing time of each project in GPU hours and not in per-node hours. In Table 1 we report the training times in hours, as well as the total power draw (in kWh) of each Jean Zay and Adastra project associated with the ALMANaCH team during 2025. We use this information to compute the total power consumption (multiplying by the PUE) of each project, also reported in Table 1.

We can further estimate the CO₂ emissions in kilograms of each single project by multiplying the total power consumption by the average CO₂ emissions per kWh in France, which were around 27g/kWh on average for 2025.²³ The total CO₂ emissions in kg for one single model can therefore be computed as:

$$\text{CO}_2e = 0.027p_t \quad (2)$$

All emissions are also reported in Table 1. The total emission estimate for the team is 20,106kg of CO₂. The carbon footprint of a single passenger on a round trip from Paris to Albuquerque (USA) (Boeing 787), flying economy, amounts to around 3,400kg of CO₂.²⁴ This means that our computing emissions from Jean Zay and Adastra for 2025 amount to under six Paris-Albuquerque trips (NAACL 2025), and therefore, given the size of our project-team, the largest source of emissions from the team are from flights to conferences, which are a necessary part of communicating about our research to the international community.

6 Highlights of the year

6.1 Awards

- Lydia Nishimwe was one of the 12 finalists of the Sorbonne Université 2025 edition of “Ma thèse en 180 secondes”.
- Nathan Godey was one of the two winners of the ATALA (learned society for French NLP) thesis prize in 2025.

¹⁸Processor page on the Intel web site.

¹⁹Cines documentation.

²⁰LUMI documentation.

²¹More information here.

²²See more information here.

²³Source: Nowtricity web site.

²⁴co2.myclimate.org estimates for 2025.

- Oriane Nédey was one of the two winners of the **best paper award at RECITAL 2025** for her article entitled “La traduction automatique dialectale: état de l’art et étude préliminaire sur le continuum dialectal de l’occitan”
- Célia Nouri and her supervisors Chloé Clavel and Jean-Philippe Cointet’s paper entitled “Graphically Speaking: Unmasking Abuse in Social Media with Conversation Insights” was selected as one of the **SAC Highlights** (Senior Area Chair Highlights) at ACL 2025 in Vienna. This designation represents about 4% of presented papers at ACL.
- Lucie Chenain won the PhD Thesis Prize at the 2nd edition of the “Avenir de la Recherche Clinique” award, presented during the 13th *Journée de la Recherche Clinique* organised by AFCROs.

7 Latest software developments, platforms, open data

7.1 Latest software developments

7.1.1 HTR-United

Keywords: HTR, OCR

Functional Description: HTR-United is a Github organization without any other form of legal personality. It aims at gathering HTR/OCR transcriptions of all periods and styles of writing, mostly but not exclusively in French. It was born from the mere necessity for projects- to possess potentiel ground truth to rapidly train models on smaller corpora.

Datasets shared or referenced with HTR-United must, at minimum, take the form of: (i) an ensemble of ALTO XML and/or PAGE XML files containing either only information on the segmentation, either the segmentation and the corresponding transcription, (ii) an ensemble of corresponding images. They can be shared in the form of a simple permalink to ressources hosted somewhere else, or can be the contact information necessary to request access to the images. It must be possible to recompose the link between the XML files and the image without any intermediary process, (iii) a documentation on the transcription practices followed for the segmentation and the transcription. In the cases of a Github repository, this documentation must be summarized in the README.

A corpus can be sub-divided into smaller ensembles if it seems necessary.

Release Contributions: First version.

URL: <https://htr-united.github.io/>

Contact: Alix Chague

7.1.2 CamemBERT-bio

Keywords: Language model, Deep learning, NLP, Transformer

Functional Description: CamemBERT-bio is a state-of-the-art french biomedical language model built using continual-pretraining from camembert-base. It was trained on a french public biomedical corpus of 413M words containing scientific documents, drug leaflets and clinical cases extrated from theses and articles. It shows significant improvement on multiple biomedical named entity recognition tasks compared to camembert-base.

URL: <http://camembert-bio-model.fr/>

Contact: Rian Touchent

7.1.3 RoCS-MT

Name: Robust Challenge Set for Machine Translation

Keywords: Machine translation, NLP, Evaluation, Robustness, User-generated content

Functional Description: RoCS-MT, a Robust Challenge Set for Machine Translation (MT), is designed to test MT systems' ability to translate user-generated content (UGC) that displays non-standard characteristics, such as spelling errors, devowelling, acronymisation, etc. RoCS-MT is composed of English comments from Reddit, selected for their non-standard nature, which have been manually normalised and professionally translated into five languages: French, German, Czech, Ukrainian and Russian. The challenge set was included as a test suite at the WMT 2023 conference. This repository therefore also includes automatic translations from the submissions to the general MT task.

URL: <https://github.com/rbawden/RoCS-MT>

Publications: [hal-04300824](#), [hal-05344725](#)

Contact: Rachel Bawden

Participants: Rachel Bawden, Benoit Sagot

7.1.4 3MT_French Dataset

Name: 3 Minutes Thesis Corpus

Keywords: Multimodal Corpus, Video annotation

Functional Description: This new resource will be useful to computer science and social science researchers working on public speaking assessment and training. It will help refine the analysis of speaking from a fresh perspective based on social-cognitive theories rarely studied in this context, such as first impressions and theories of primacy and recency.

URL: <https://zenodo.org/records/7603511#.Y90413CZ0Uk>

Publication: [hal-04366763](#)

Contact: Chloe Clavel

7.1.5 CATMuS Medieval (Model)

Name: Consistent Approach to Transcribing Manuscripts - Medieval model

Keyword: Handwritten Text Recognition

Functional Description: CATMuS (Consistent Approach to Transcribing Manuscript) Medieval is a model for automatically transcribing medieval manuscripts using Latin scripts, in particular Old and Middle French, Latin, Spanish (and other languages of Spain), and Italian. The model was trained on the largest and most diverse dataset known for medieval manuscripts in Latin scripts, with more than 110 000 lines of training data.

Contact: Thibault Clerice

Partners: University of Toronto, Ecole nationale des chartes, CIHAM UMR 5648, VeDPH - Ca' Foscari, Université de Genève, ENS Lyon

7.1.6 HTRomance

Keyword: Handwritten Text Recognition

Functional Description: The ground truth produced as part of the HTRomance project aims to provide diverse data, from the 12th century to the 19th century, for training handwritten text recognition models. It covers the following languages: Latin, various states of French, Spanish, Occitan and Italian.

URL: <https://htromance-project.github.io/>

Contact: Thibault Clerice

Partners: VeDPH - Ca' Foscari, Ecole nationale des chartes, CIHAM UMR 5648, ENS Lyon

7.1.7 eScriptorium Documentation

Name: Open documentation for eScriptorium

Functional Description: Collaborative and open documentation redacted using the functionalities offered by Github et deployed thanks to Readthedocs. It offers an illustrated description of all the features of the eScriptorium application, which does not offer any other form of complete documentation, as well as tutorials.

URL: <https://escriptorium.readthedocs.io/>

Contact: Alix Chague

7.1.8 CATMuS Medieval (Dataset)

Name: Consistent Approaches to Transcribing Manuscripts - Medieval Dataset

Keywords: Handwritten Text Recognition, HTR, OCR

Functional Description: Developed through collaboration among various institutions and projects, CAT-MuS provides an inter-compatible handwritten text recognition dataset spanning more than 240 manuscripts and incunabula in 10 different languages, comprising over 170,000 lines of text and 5 million characters spanning from the 8th century to the 16th.

Release Contributions: - 40 new manuscripts

Publication: [hal-04453952](https://hal.archives-ouvertes.fr/hal-04453952)

Contact: Thibault Clerice

Partners: CIHAM UMR 5648, Ecole nationale des chartes, University of Toronto, Antwerp University

7.1.9 Counter dataset

Name: Counter Radicalization Dataset

Keywords: NLP, Multilingual corpus, Data protection, Pseudonymization, Online radicalization

Functional Description: This dataset includes multilingual content from forums, Telegram, social media, and other sources in English, French, and Arabic. It covers various radical ideologies and is pseudonymized to protect privacy while maintaining data utility. It includes annotations for call to action, radicalization level, and named entity recognition.

URL: <https://gitlab.inria.fr/ariabi/counter-dataset-public>

Contact: Djame Seddah

7.1.10 CamemBERTav2

Keywords: Language model, French

Functional Description: It is the second version of the CamemBERTa model, which is based on the DeBERTaV2 architecture with the Replaced Token Detection (RTD) objective.

The new update includes: 1) Much larger pretraining dataset: 275B unique tokens (previously 32B).
2) A newly built tokenizer based on WordPiece with 32,768 tokens, addition of the newline and tab characters, support emojis, and better handling of numbers (numbers are split into two digits tokens).
3) Extended context window of 1024 tokens

Contact: Benoit Sagot

7.1.11 CamemBERTv2

Keywords: Language model, French

Functional Description: CamemBERTv2 is a French language model pretrained on a large up-to-date corpus of 275B tokens of French text. The model still based on the BERT architecture, demonstrates improved performance across a variety of NLP tasks, over the previous version

URL: <https://huggingface.co/almanach/camembertv2-base>

Contact: Benoit Sagot

7.1.12 LADaS

Name: Layout Analysis Dataset with Segmonto

Keyword: Layout Analysis

Functional Description: LADaS is a dataset for training layout analysis on documents from the 16th to the 21st century. It helps reproduce documents in XML TEI.

Publication: [hal-04513725](https://hal.archives-ouvertes.fr/hal-04513725)

Contact: Thibault Clerice

7.1.13 CamemBERT-bio-gliner

Keywords: Language model, Deep learning, NLP, Transformer

Functional Description: CamemBERT-bio-gliner is a Named Entity Recognition (NER) model capable of identifying any french biomedical entity type using a BERT-like encoder. It provides a practical alternative to traditional NER models, which are limited to predefined entities, and Large Language Models (LLMs) that, despite their flexibility, are costly and large for resource-constrained scenarios. CamemBERT-bio is used as a backbone.

Contact: Rian Touchent

7.1.14 CoMMuTE

Name: Contrastive multilingual and multimodal translation evaluation

Keywords: Machine translation, Evaluation, Image analysis

Functional Description: CoMMuTE is a contrastive evaluation dataset designed to assess the ability of multimodal machine translation models to exploit images in order to disambiguate the sentence to be translated. In other words, given a sentence containing a word that can be translated in several ways, the additional image determines the meaning of the word to be translated. The model must then take the image into account to propose a correct translation. CoMMuTE is available from English into French, German and Czech.

URL: <https://github.com/MatthieuFP/CoMMuTE>

Publications: [hal-03977982](#), [hal-04736377](#)

Contact: Matthieu Futral-Peter

Participants: Matthieu Futral-Peter, Rachel Bawden, Benoit Sagot

7.1.15 mOSCAR

Keywords: Raw corpus, Multilingual corpus, Multimodal Corpus, Multimodality, Text-image processing, Web crawling

Functional Description: mOSCAR is the first large-scale multilingual and multimodal web-crawled document corpus. It covers 163 languages, 315M documents, 214B tokens and 1.2B images. We carefully conduct a set of filtering and evaluation steps to make sure mOSCAR is sufficiently safe, diverse and of good quality.

URL: <https://oscar-project.github.io/documentation/versions/mOSCAR/>

Publication: [hal-04629451](#)

Contact: Matthieu Futral-Peter

Participants: Matthieu Futral-Peter, Benoit Sagot, Rachel Bawden, Julien Abadji, Armel Zebaze Dongmo, Cordelia Schmid, Remi Lacroix

7.1.16 Concordancer

Name: Concordancer

Keywords: Natural language processing, Document analysis, Terminological analysis

Scientific Description: Concordancer enables the detection in a parsed corpus of a set of terms from a terminology. It also detects potential variants of these terms by applying various transformations (such as reduction, acronym, morphological derivations, synonymy, etc.), in order to study variations in the use of a term within a same document.

Functional Description: Concordancer enables the detection in a parsed corpus of a set of terms from a terminology. It also detects potential variants of these terms by applying various transformations (such as reduction, acronym, morphological derivations, synonymy, etc.), in order to study variations in the use of a term within a same document.

Contact: Éric De La Clergerie

7.1.17 SWELLS

Name: Specialized Workbench for the Explicit Learning of Linguistic Structures

Keywords: Machine translation, Natural language processing, Low resources languages, Large Language Models, Language workbench

Functional Description: SWELLS is a framework specially designed to study explicit learning in Large Language Models (LLMs) by means of machine translation experiments involving constructed languages (conlangs). It provides datasets and scripts that enable a controlled assessment of an LLM's ability to learn and apply metalinguistic rules presented in grammar books.

URL: <https://github.com/mmarmonier/SWELLS>

Publication: [hal-04991098](#)

Contact: Malik Marmonier

Participants: Malik Marmonier, Rachel Bawden, Benoit Sagot

7.1.18 ACReFOSC

Name: A Companion Repository to the French OLDI Seed Corpus

Keyword: Machine translation

Functional Description: ACReFOSC provides the code and data needed to automatically generate preference datasets for machine translation, supporting research in preference optimization and the post-training of neural models and LLMs for machine translation purposes.

URL: <https://github.com/mmarmonier/ACReFOSC>

Publication: [hal-05375157](https://hal.archives-ouvertes.fr/hal-05375157)

Contact: Malik Marmonier

Participants: Malik Marmonier, Rachel Bawden, Benoit Sagot

7.1.19 CoMMA

Name: Corpus of Multilingual Medieval Archives

Keywords: Historical language, Raw corpus, TEI, XML, HTR, OCR

Functional Description: CoMMA is a large-scale corpus of medieval manuscripts produced through automatic text recognition. The corpus contains around 3.3b tokens drawn from more than 32,700 digitized manuscripts in Latin and Old French.

Contact: Thibault Clerice

Partners: Université de Genève, CIHAM UMR 5648

7.1.20 NeWMe

Name: The NeWMe Corpus (Negotiating Word Meaning)

Keywords: Raw corpus, Natural language, Dialogue, Semantics, Discourse

Functional Description: The NeWMe corpus contains 661 annotated examples of Word Meaning Negotiation (WMN: sequences in conversation where speakers discuss word meaning) as well as related conversational phenomena. Annotations contain the type of phenomenon as well as parts of the WMN: a trigger (problematic word usage), indicator (signaling a problem with word meaning), and the negotiation (a meta-linguistic discussion of word meaning). NeWMe contains oral conversations from Switchboard and the BNC as well as written online debates from Reddit.

Contact: Aina Gari Soler

7.1.21 SPOT

Keyword: Social networks

Functional Description: SPOT (Stopping Points in Online Threads) is the first annotated corpus operationalizing the sociological concept of stopping point (i.e critical interventions) as a reproducible task for computational study. The SPOT dataset comprises approximately 43,000 comments from French Facebook discussions linked to URLs flagged as false information, enriched with contextual metadata (parent post, parent comment, page/group and source features)

Contact: Celia Nouri

7.1.22 eScriptorium

Keywords: HTR, OCR, Web Application, Annotation tool

Functional Description: Web application for manual, semi-automatic, and automatic segmentation and transcription of printed or handwritten text documents, with the possibility of training or reusing transcription models.

News of the Year: Thibault Cl rice joined the boards of directors for the development of eScriptorium. Hassen Aguilı became the main maintainer of the software in december 2025.

Contact: Thibault Clerice

Participant: Hassen Aguilı

7.1.23 ocDI

Keywords: Supervised models, Variety identification, Natural language processing

Functional Description: A series of models for the identification of the Occitan dialect of a text. The models were trained on several dialect-labelled Occitan datasets to predict a dialect out of 8 labels corresponding to the following Occitan varieties: Languedocian, Proven al, Limousine, Auvergnat, Vivaro-alpine, Ni ard, Gascon, and Aranese.

Three variants of the model are available to correspond to various usages and constraints: - ocDI-SVM: a linear classifier based on full words and n-grams of characters, that enables easy interpretation of the model predictions, and obtained the best recall for Auvergnat, Gascon, Ni ard and Vivaro-alpine. - ocDI-FastText: a very fast classifier based on word and character n-gram static embeddings, which obtained the best recall for Gascon and Limousine. - ocDI-BERT: a fine-tuned BERT model that obtained the best overall performance

For data mining purposes, it is recommended to use the model with the best recall on the target variety as our models tend to over-predict the most represented classes (esp. Languedocian).

Contact: Oriane Nedey

Participants: Oriane Nedey, Thibault Clerice, Rachel Bawden, Benoit Sagot

7.1.24 OcWikiDialects

Keywords: Raw corpus, Low resources languages

Functional Description: OcWikiDialects is a corpus of over 7k articles from the Occitan Wikipedia, with rich metadata including dialect labels for 8 varieties of Occitan (Proven al, Languedocian, Gascon, Auvergnat, Limousine, Vivaro-alpine, Ni ard, Aranese) and 2 transitional varieties (Marchese and Aguianese). It contains full articles in Markdown and WikiText formats, and segmentations into paragraphs and sentences, for a total of approximately 4M tokens.

Contact: Oriane Nedey

7.1.25 GAPeron

Name: GAPeron

Keywords: LLM, LLM Safety

Functional Description: We release Gaperon, a fully open suite of French-English-coding language models designed to advance transparency and reproducibility in large-scale model training. The Gaperon family includes 1.5B, 8B, and 24B parameter models trained on 2-4 trillion tokens, released with all elements of the training pipeline: French and English datasets filtered with a neural quality classifier, an efficient data curation and training framework, and hundreds of intermediate checkpoints. Through this

work, we study how data filtering and contamination interact to shape both benchmark and generative performance. We find that filtering for linguistic quality enhances text fluency and coherence but yields subpar benchmark results, and that late deliberate contamination – continuing training on data mixes that include test sets – recovers competitive scores while only reasonably harming generation quality. We discuss how usual neural filtering can unintentionally amplify benchmark leakage. To support further research, we also introduce harmless data poisoning during pretraining, providing a realistic testbed for safety studies. By openly releasing all models, datasets, code, and checkpoints, Gaperon establishes a reproducible foundation for exploring the trade-offs between data curation, evaluation, safety, and openness in multilingual language model development.

URL: <https://huggingface.co/collections/almanach/gaperon>

Publication: [hal-05410121](https://hal.archives-ouvertes.fr/hal-05410121)

Contact: Djame Seddah

Participants: Nathan Godey, Wissam Antoun, Rian Touchent, Djame Seddah, Éric De La Clergerie, Benoit Sagot, Rachel Bawden

7.1.26 ModernCamemBERT

Keyword: LLM

Functional Description: The goal of ModernCamemBERT was to run a controlled study by pretraining ModernBERT on the same dataset as CamemBERTaV2, a DeBERTaV3 French model, isolating the effect of model design. Our results show that the previous model generation remains superior in sample efficiency and overall benchmark performance, with ModernBERT’s primary advantage being faster training and inference speed. However, the new proposed model still provides meaningful architectural improvements compared to earlier models such as the BERT and RoBERTa CamemBERT/v2 model.

URL: <https://huggingface.co/collections/almanach/moderncamembert>

Contact: Wissam Antoun

7.1.27 Glyphea

Keywords: Web Application, Cryptography, HTR

Functional Description: Web application for transcribing encrypted handwritten texts

Contact: Cécile Pierrot

7.2 New platforms

Participants: Thibault Clerice, Hassen Aguilu, Benjamin Kiessling.

- **Textile** (MIT): A web interface for browsing and searching the CoMMA (Corpus of Medieval Manuscript Archives) collection. Textile provides a way to explore medieval manuscripts with full-text search, document viewing, and multilingual support.
- **Textile Backend** (MIT): The Textile backend is a web API that supports very large corpora based on Dapytains and provides a full-text search option.
- **Glyphea** (Apache 2.0): Web application for transcribing encrypted handwritten texts

7.3 Open data

- **OSCAR** (CC-BY): Huge multilingual web-based corpus
- **Counter dataset** (Full data set only available for request for research only. Annotations examples available on the dataset gitlab’s repo can be freely released (CC-BY-NC-SA)): An open-source pseudonymized dataset aimed at facilitating research on radicalization detection with NER annotations. It is the first publicly available multilingual dataset for radicalization detection, gathered from diverse social media sources.
- **CATMuS Medieval (Dataset)** (CC-BY): Large-scale diverse dataset for handwritten text recognition of medieval manuscripts
- **mOSCAR** (CC-BY): Large-scale multilingual, multimodal (text-image) web-crawled corpus
- **LADaS** (CC-BY): LADaS (Layout Analysis Dataset with SegmOnto) is a diachronic diagenetic layout analysis dataset (16th-21st c.)
- **SWELLS** (MIT (code), CC BY-SA 4.0 (datasets)): SWELLS makes it possible to assess, in a controlled way, the ability of language models to assimilate specific aspects of an unknown language on the basis of grammar book excerpts added to their prompts.
- **ACReFOSC** (CC BY-SA 4.0): Generates fine-tuning datasets for preference optimization in machine translation.
- **CoMMA** (CC BY- 4.0): CoMMA is a large-scale corpus of medieval manuscripts produced through automatic text recognition. The corpus contains around 3.3b tokens drawn from more than 32,700 digitized manuscripts in Latin and Old French.
- **NeWMe**: A corpus of annotated instances of Word Meaning Negotiation (sequences in conversation where speakers discuss word meaning) from existing oral and written conversational corpora.
- **SPOT** (CC-BY 4.0): SPOT (Stopping Points in Online Threads) is a French corpus of 43k Facebook comments annotated for the presence of stopping points (critical interventions)
- **CoMMuTE** (CC-BY-SA-4.0): A contrastive evaluation dataset for multimodal (text-image) machine translation.
- **RoCS-MT** (CC-BY-NC): Robust Challenge Set for Machine Translation
- **OcWikiDialects** (CC-BY): OcWikiDialects is a corpus derived from Occitan Wikipedia, featuring diverse metadata, including dialect annotations.

8 New results

8.1 Language Modelling

Participants: Benoît Sagot, Rachel Bawden, Djamé Seddah, Éric Villemonte De La Clergerie, Wissam Antoun, Pierre Chambon, Matthieu Futeral-Peter, Nathan Godey, Rian Touchent, Armel Zebaze Dongmo.

8.1.1 Language Models

Our research activity on language modelling in 2025 has focused on understanding, improving, and demystifying large-scale transformer models along three complementary axes: architectural evaluation under controlled conditions, openness and data-centric analysis in large language model (LLM) training, and efficiency at inference time, particularly for long-context generation.

A first line of work addresses a growing challenge in the field: disentangling the respective contributions of model architecture and pretraining data to downstream performance. Recent transformer variants often report gains over previous models, but these claims are frequently confounded by undisclosed data or incomparable training setups. To address this, in the context of Wissam Antoun’s PhD thesis supervised by Djamé Seddah and Benoît Sagot, we conducted a controlled comparison between two recent encoder architectures by pretraining ModernBERT on exactly the same data as CamemBERTaV2, a DeBERTaV3-based French model [24]. This study shows that, when data is held constant, the previous model generation remains superior in both sample efficiency and overall benchmark performance. ModernBERT’s main advantages lie instead in engineering-oriented improvements, such as support for longer contexts and faster training and inference. Beyond this comparison, we also show that higher-quality pretraining data mainly accelerates convergence without substantially improving final performance, suggesting a degree of saturation in commonly used benchmarks. Overall, this work advocates for more rigorous evaluation protocols and highlights that architectural novelty should be assessed independently from data scale and quality.

A second major contribution is our commitment to transparency and reproducibility in large-scale language model training, embodied in the release of the Gaperon model suite [76], a collaborative effort primarily supervised by Djamé Seddah and Benoît Sagot and involving Nathan Godey, Wissam Antoun, Rian Touchent, Éric Villemonte De La Clergerie and Rachel Bawden. Gaperon consists of fully open French-English-code autoregressive models ranging from 1.5B to 24B parameters, trained on trillions of tokens. Crucially, we release not only the final models, but also the datasets, filtering tools, training code, and hundreds of intermediate checkpoints. This openness enables a systematic investigation of how data filtering, contamination, and evaluation practices interact. Our findings reveal a tension between linguistic quality and benchmark performance: aggressive neural filtering improves fluency and coherence but degrades benchmark scores, while deliberate late-stage contamination can recover competitive results at a moderate cost to generation quality. These results shed light on how standard evaluation pipelines may inadvertently encourage benchmark leakage. By additionally introducing controlled, harmless data poisoning, Gaperon provides a realistic testbed for future work on robustness and safety, while establishing a solid empirical foundation for data-centric LLM research.

Finally, in the context of Nathan Godey’s PhD, supervised by Benoît Sagot and Éric Villemonte De La Clergerie, we investigated efficiency challenges at inference time, focusing on the memory bottleneck induced by the Key-Value (KV) Cache in autoregressive models with long contexts [77], in collaboration with researchers from the University of Edinburgh, the University of Rome and Miniml.AI. We introduce Q-Filters, a training-free KV cache compression method based on previously unexplored properties of query and key vectors. By filtering less relevant key-value pairs using a simple, context-agnostic projection, Q-Filters achieves strong compression while remaining compatible with optimised attention implementations such as FlashAttention. Experiments show that it matches or outperforms existing methods across both retrieval and generation tasks, enabling extreme compression levels with minimal performance degradation.

8.1.2 Multimodal Language Modelling

In 2025 we continued our activities in multimodal language modelling in two different directions. A first research axis, in collaboration with META through two CIFRE PhDs defended in 2024 under Benoît Sagot’s academic supervision, investigated how to tightly integrate language and speech within a single generative framework. As a final outcome of this collaboration, we introduced SPIRIT-LM, a multimodal language model that treats text and speech as a unified token stream [19]. Starting from a 7B-parameter pretrained text language model, we continuously train on mixed sequences of text and discretised speech units, interleaved at the word level. This design enables the model to seamlessly switch between modalities, combining the semantic reasoning strengths of text language models with the expressive richness of speech. Beyond a base version relying on phonetic units, an expressive variant incorporates pitch and style information, allowing the model to capture prosody and expressivity. SPIRIT-LM demonstrates strong few-shot cross-modal generalisation, supporting tasks such as ASR, TTS and speech classification within a single architecture.

A second, complementary research axis addresses the data foundations required for scalable and inclusive text-image multimodal learning, in the context of Matthieu Futral-Peter’s PhD thesis, defended in 2025, and co-supervised by Rachel Bawden and Benoît Sagot for ALMAnaCH and by Cordelia Schmid for WILLOW (another research team from the Inria Paris research centre, specialised in computer vision). While prior work has shown how combining text and images can be useful, including our own work on the use of the

image modality to improve MT [107], existing multimodal (text-image) training corpora are either private or limited to English, and generally limited to captioning data. To overcome this bottleneck, we developed and released mOSCAR, the first large-scale multilingual and multimodal document corpus crawled from the web [33]. mOSCAR spans 163 languages and combines text with over a billion images, enabling research on multimodal models for a truly global set of languages. Through extensive filtering and evaluation, we ensure data quality and safety. To prove mOSCAR’s utility, we train a multilingual OpenFlamingo from a Gemma-2B language model on a subset of mOSCAR and captioning data derived from the LAION-400M corpus and translated using NLLB. We compare against a similar model trained on captioning data only and show we obtain a strong boost in few-shot learning, confirming previous findings for English. Further results specifically for translation are mentioned in the dedicated section (Section 8.3.5).

8.1.3 Code Generation

In 2025 we continued our collaboration with META on code generation with LLMs, in the context of Pierre Chambon’s CIFRE PhD, under Benoît Sagot’s academic supervision. Our work this year has focused on advancing the evaluation and training of language models for code generation, with particular emphasis on algorithmic reasoning and computational constraints. We developed and released BigO(Bench) [72], a large-scale benchmark specifically designed to assess whether generative models can understand and produce code that satisfies explicit time and space complexity requirements. Unlike prior evaluations that largely ignore complexity, BigO(Bench) combines a principled profiling-based framework to infer algorithmic complexity of Python programs with a rich dataset of more than 3,000 problems and over one million annotated solutions derived from competitive programming. Experimental results reveal a nuanced picture: while modern token-space reasoning models excel at producing functionally correct code, they often struggle to reliably meet complexity constraints, suggesting limited generalisation beyond training-time rewards.

Building on these findings, we have initiated work on reinforcement learning methods with the aim of improving code generation quality and robustness under diverse objectives. In particular, we are exploring post-training with Group Relative Policy Optimisation (GRPO) [123], an efficient RL approach that replaces value models with Monte Carlo estimates. Our ongoing work investigates GRPO in a multi-task setting, where tasks are stochastically sampled and training is organised through a global pool of workers and trainers. This line of research lays the groundwork for more stable, scalable, and flexible RL-based optimisation of code-generating language models.

8.2 Social, Cultural and Political Computing

Participants: Chloé Clavel, Alisa Barkar, Yanzhu Guo, Gabrielle Le Bellier, Cecilia Graiff, Carlo Santagiustina.

8.2.1 LLMs’ Ability to Understand the Concept of Persuasiveness

Alisa Barkar completed her PhD in December 2025. She was co-supervised by Chloé Clavel, Mathieu Chollet (associate Professor at the University of Glasgow) and Beatrice Biancardi (Associate Professor at CESI). During her last year, we investigate the application of LLMs for assessing public speaking (PS) by predicting persuasiveness. We propose a novel framework where LLMs evaluate criteria derived from educational literature and feedback from PS coaches, offering new interpretable textual features. We demonstrate that persuasiveness predictions of a regression model with the new features achieve a Root Mean Squared Error (RMSE) of 0.6, underperforming approach with hand-crafted lexical features (RMSE 0.51) and outperforming direct zero-shot LLM persuasiveness predictions (RMSE of 0.8). Furthermore, we find that only LLM-evaluated criteria of language level is predictable from lexical features (F1-score of 0.56), disapproving relations between these features. Based on our findings, we criticise the abilities of LLMs to analyse PS accurately. To ensure reproducibility and adaptability to emerging models, all source code and materials are publicly available on GitHub. This work was published at an international conference (ICAART) [25] and a national conference (TALN) [53].

8.2.2 Evaluating Linguistic Diversity of LLM outputs

Yanzhu Guo completed her PhD on automatic text evaluation in 2025 under the supervision of Chloé Clavel. Building on her earlier work on diversity decline when recursively training language models on generated text [108], we propose a comprehensive framework for evaluating LLMs from multiple linguistic diversity perspectives [17]. Using this framework, we benchmark six state-of-the-art LLMs across lexical, syntactic, and semantic diversity on five NLG tasks, with an in-depth case study on syntactic diversity. We further analyse how model scale, pretraining data, instruction tuning, decoding strategies, and quantisation affect diversity. Our results show that, despite strong task-solving abilities, current LLMs fall short of human linguistic richness, particularly for creative tasks such as story generation. Notably, instruction tuning increases lexical diversity but constrains syntactic and semantic diversity, suggesting reduced expressive flexibility. These findings highlight a risk of homogenisation in LLM-generated content and highlights the need to prioritise linguistic diversity alongside traditional performance metrics.

8.2.3 Political Analysis

In this research direction, led by Carlo Santagiustina, we provide large-scale comparative evidence on how citizens across the European Union mobilise for e-petitions through Social Media, and how such mobilisation is shaped by ideological orientations toward political elites and economic policy [87]. By analysing over 1.8 million X/Twitter profiles across five EU member states (Germany, France, Italy, the Netherlands, and Poland), we model the likelihood that users share petition-related content as a function of their individual ideological positions, the average ideological positions of the parties they follow, the distance between the two, and a set of pseudo-demographic and user-activity controls. Our findings show that online petition mobilisation in Europe is primarily driven by individual-level ideological attitudes, rather than by the ideological orientation of parties or political groups as such. A key result is that mobilisation for e-petitions often reflects perceived gaps in representation rather than coordinated party-led engagement. While the average ideological positions of parties followed by users do not significantly increase the likelihood of petition sharing, ideological distance from those party positions (especially along the anti-elite dimension) substantially raises the probability of mobilisation. This indicates e-petitions tend to function as channels for citizens to express political discontent when they feel insufficiently represented by actors they otherwise follow or identify with, highlighting their role as indicators of representational strain within European democracies. Our work highlights the strategic value of analysing Social Media-based mobilisation as a complementary tool for democratic governance at national and EU levels. Patterns of petition sharing, especially when concentrated in particular ideological groups, can serve as early signals of rising political discontent not yet visible through electoral outcomes or traditional surveys. Monitoring e-petition mobilisation on Social Media can help anticipate contention, improve responsiveness to citizens' needs and concerns, and better align institutional engagement with how Europeans express political preferences and dissatisfaction online.

8.2.4 Cultural Adaptation of LLMs

In the context of Gabrielle Le Bellier's PhD, supervised by Chloé Clavel and Benoît Sagot, we explored the lack of cultural commonsense knowledge of LLMs. LLMs struggle to answer cultural questions about daily life, such as "What is a common snack for preschool kids in [country]?", where we focus on a culture at a country scale. They show disparate performances between high-resource cultures and low-resource ones, for instance with high accuracies for the United Kingdom and low accuracies for North Korea or Ethiopia. By using soft prompt methods on OLMo-7B-Instruct, we showed that the model's ability to answer cultural commonsense questions can be improved. These methods require few computational resources by tuning only a few parameters added to the frozen model (approximately 0.001% for prompt-tuning and 1% for prefix-tuning). We fine-tuned the soft prompts on cultural multi-choice question answering and obtained one soft prompt per country. We observed cultural proximity by evaluating a prompt on other cultures' questions. For instance, the Spain prompt-tuned model is better at answering questions about Mexico than questions about North Korea. Moreover, we argue that we can interpret similarities between soft prompts to deduce cultural proximity. This work was co-funded by Benoît Sagot's PRAIRIE chair position and Chloé Clavel's PRME SINNet ANR project, in collaboration with the "Action exploratoire" SaLM.

8.2.5 Argumentation in Political Debates

In the context of Cecilia Graiff’s PhD, funded by Benoît Sagot’s PRAIRIE chair position and co-supervised by Benoît Sagot and Chloé Clavel in collaboration with the “Action Exploratoire” SaLM (Socially Aware Language Models), we benchmarked the performance of language models on cross-lingual and cross-country argument component detection, focusing on political data from the US and France. We considered natural language arguments as composed by claims and premises, following the most reliable approaches in argument mining. To overcome the lack of data in this field, we introduced FrenchPolArg, a corpus of argumentative political discourse in French, and we automatically translated and projected the annotations of ElecDeb60to20, a corpus of US presidential debates annotated for argument components and relations. We benchmarked three different cross-lingual and cross-country pipelines, and we compared their results to find the best-performing one, namely model transfer. We obtained promising results to be integrated in semi-automatic annotation workflows to reduce the time and cost of annotations. Moreover, we are launching an annotation campaign with expert annotators from SciencesPo Paris, in order to obtain a ground truth and continue our experiments. In the future, this PhD aims to explore the abilities of language models to understand reasoning patterns.

8.3 Machine Translation (MT)

Participants: Rachel Bawden, Benoît Sagot, Thibault Clérice, Lydia Nishimwe, Matthieu Futeral, Armel Zebaze, Nicolas Dahan, Ziqian Peng, Oriane Nédey, Malik Marmonier, Panagiotis Tsolakis, Rasul Jasir Dent.

8.3.1 MT Benchmarking and Shared Tasks

As in previous years, Rachel Bawden was a member of the organising committee of the general MT shared task at the main conference in MT (WMT) [82, 40] (she has previously also been part of the committee for the biomedical task, but which did not take place this year). The purpose of this shared task is the evaluation of models submitted by participants on the automatic translation of texts from multiple domains, namely news, literary, speech and social. The shared task has evolved over time, and this year the focus was particularly on document-level MT. Rachel Bawden assisted notably with the social media data, which this year included screenshots of threads from Mastodon in order to enable the participation of multimedial text-image models. Rachel Bawden and Benoît Sagot also participated in the test suite shared task, with an updated version of the RoCS-MT (Robust Challenge Set for MT) challenge set [27], a parallel set designed for the evaluation of MT of highly non-standard user-generated content (UGC), which is detailed more in Section 8.3.3 below. Rasul Jasir Dent was also member of the organising committee for the first shared task in Creole MT, at the same conference, which encouraged submissions of MT systems and datasets for MT involving Creole languages.

In terms of benchmarking, Rachel Bawden continued her participation in an initiative led by Jesujoba Alabi (a previous engineer in the team and now a PhD student at Saarland University) on the creation of a document-level domain-specific (health and tech) corpus for African languages, AfriDoc-MT. The benchmark was used to evaluate LLM performance at the document level, showing that document-level MT remains challenging for these languages. The work was finalised and published at EMNLP 2025 [22].

8.3.2 MT for Open Science

We continued our work on MT for scientific documents in the context of the MaTOS (Machine Translation for Open Science) ANR project [26], led by François Yvon (CNRS). There is a specific focus on English–French and French–English translation, since the high quality of MT for these directions allows us to work on more high-level challenges such as term translations, consistency and document coherence. Two PhD students are currently co-supervised by Rachel Bawden jointly with François Yvon in the project: Ziqian Peng (recruited by the CNRS) on document-level MT for scientific documents and Nicolas Dahan on the evaluation of MT.

Following her work on improving translation for longer documents last year (with full document context), this year Ziqian Peng has mainly been focusing on extracting relevant local and distant contexts to help translation. The main motivation for the work is that context can be useful, but is typically sparse (not all context will be useful) and it can be inefficient and sometimes detrimental to include full document context

when translating. It may be beneficial to be able to extract context from the left context in a document (i.e. the previously translated sentences) that can either be local or more distant, without having to include full document context. With the proposed self-RAMT approach (for retrieval-augmented MT framework), it is possible to dynamically extract contexts from the source text and the associated translation. We carried out experiments on three LLMs, considering three different criteria to select relevant contexts to include: (i) cosine similarity using LaBSE [105], a variant of BM25L [110] and PMI (to identify contexts that are most likely to make a difference). We carry out experiments on TED talks as well as parallel scientific articles, considering three translation directions. Our results show that integrating distant contexts does improve translation quality as measured by reference-based scores and consistency metrics.

In addition to models fine-tuned for the project (on scientific documents from Natural Language Processing (NLP) and Earth and Planetary Sciences), described in [86], we have also been contributing to the collection and processing of additional resources for training and evaluation of MT models for scientific documents. In a paper currently under submission, we describe the collection of English–French parallel corpora for NLP and Earth and Planetary Sciences, including train, development and test sets and covering 14k abstracts and 104 full-length articles. The parallel data comes from a variety of sources and includes human-produced translations as well as automatic translations to fill the gaps where no reference translations exist. We test the performance of eight MT systems on our test sets and find that fine-tuning on our data reduces the gap between open LLM-based systems and commercial ones. We also find that the performance of recent LLMs can worsen when translating full articles instead of translating them at the paragraph level.

In the context of Nicolas Dahan’s PhD, we continued to look at ways of evaluating document-level consistency and coherence in translations. We finalised the MetaDocEval dataset, comprising contrastive test examples of correct and incorrect translations. The incorrect samples were automatically generated to contain specific errors designed to stress test metrics on potential errors of document MT systems (e.g. truncating translations, shuffling sentences, changing pronouns) and to see the effect on metrics of phenomena to which we would expect them to be robust (splitting sentences, use of synonyms). Testing on English-to-French, Spanish and German, we evaluated several MT metrics, using multiple segmentation strategies and span lengths. Our experiments showed that some metrics highly sensitive to local changes under-estimate long-range disruptions, and vice versa and that certain perturbations (e.g. related to coherence and repetition) reveal shortcomings in metrics that would appear robust under sentence-level evaluations. The paper is also under submission.

We have also turned our attention to the evaluation of term use in translation, since this is a highly important aspect of scientific document translation; terms must not only be translated correctly but they must also be translated in a coherent way in the document (this can correspond to consistency in some cases, but it also corresponds to the appropriate use of terminological variants such as acronyms, synonyms, and reductions). With the aim of modelling terminological use within documents, Éric de la Clergerie has been leading research into tools to identify terms and their variants, with significant developments made to the Concordancer tool 7.1.16, described in another submitted paper. The tool uses a predefined glossary to detect terms within documents and then detects variants based on variation of the syntactic structure of terms, their morphological variants, synonyms and reductions. It also detects chains of variants within documents, including those from the glossary and those detected dynamically. The tool can then in turn be used to enrich terminologies once it has been applied to scientific documents, which we do for the NLP and Earth and Planetary Sciences fields. We also carried out preliminary experiments in the medical domain.

The Concordancer is being used in ongoing research to model terminological use in human and automatic translations as part of Nicolas Dahan’s PhD, detecting terms and their variants to identify what the differences in use are and what constitutes a correct use of a term, especially in the context of the document. The aim is to use this analysis to develop an automatic metric that can be used to evaluate term use in document-level MT.

One of the final outcomes of the project will hopefully be an online tool to allow authors of scientific articles to automatically translate their abstracts and then post-edit them when submitting papers to the [HAL open archive](#) (at least for Inria’s instance of HAL). Panagiotis Tsolakis is currently working on an implementation to do just that, involving translating abstracts with our dedicated trained models (and testing different scenarios including zero-shot, few-shot translating with example selection) and then applying multiple verification steps to ensure a reasonable quality of automatic translation.

8.3.3 MT for Non-Standard Data

MT performance for high-resource, edited texts has improved dramatically over recent years, and LLMs have also reached state-of-the-art levels. However, challenges still remain for texts that display variation with respect to standard language varieties. The case of translation for low-resource dialects is discussed in the next section (Section 8.3.4). In this section, we cover this year’s contributions in the translation of contemporary user-generated content (UGC) and in the resolution of abbreviations in Mediaeval manuscripts, framed as a translation task.

User-Generated Content (UGC) MT Texts found on social media are characterised by various phenomena not typically present in standard edited texts and which present challenges for MT (e.g. spelling mistakes, acronyms, truncations, and contractions). It is important to develop MT models that are robust, which means they are able to translate these kinds of text just as well as if the texts had not displayed non-standard variation.

2025 marked the finalisation of Lydia Nishimwe’s PhD on robust MT, supervised by Rachel Bawden and Benoît Sagot [68], defended in June. Following her previous work on robust multilingual sentence embedding space, RoLASER, trained using using distillation and synthetic non-standard data [115], we tested the extension of the approach to the SONAR embedding space [103]. The main advantage of SONAR for translation is the presence of decoders in addition to the encoder, which enable text generation from the space and enabled us to carry out experiments on the translation of UGC. Similar results were obtained for the original SONAR and for the new embedded space trained to be robust, showing that the original space was already robust. We also found that increased robustness in the encoder does not necessarily lead to clear improvements in translation performance, and that standard encoder-decoder MT models trained on in-domain data still outperform such models, showing that domain mismatch also plays a crucial role.

In terms of evaluation, we also explored theoretical questions related to the evaluation of UGC translation, since what is considered a ‘good’ translation depends how standard the output is expected to be [85]. We studied the translation guidelines used for multiple UGC test sets through a comparison of how each handles 12 selected non-standard phenomena, revealing fundamental differences in the expected standardness of outputs. We then evaluated multiple LLMs, looking at their default translation behaviour and studying their sensitivity to explicit translation guidelines, aligning (or not) with the guidelines of the test sets. We called for better acknowledgment of fair evaluation with respect to translation guidelines and better documentation during dataset creation. This work is expected to be finalised and submitted in 2026.

As mentioned in Section 8.3.1, we also produced a new version of RoCS-MT (Robust Challenge Set for Machine Translation) [27]. This new version includes (i) minor corrections of normalisation, (ii) corrections to reference translations and the addition of alternative references to accommodate for different possible genders (e.g. of speakers) and (iii) a redesign and re-annotation of normalisation spans for further analysis of different non-standard UGC phenomena. We hope to expand how the dataset can be used for evaluation in 2026, including by improving analysis tools and allowing for easier comparison of system outputs.

Resolution of abbreviations in Mediaeval manuscripts As part of the PaRAMHTRS project [94], financed by the BNF Datalab, led by Thibault Clérice and in collaboration with David Smith from Northeastern University in Boston, we have begun exploring approaches to the resolution of abbreviations in Mediaeval Latin and Old French manuscripts. The first part of the project involved the production of parallel data for training and evaluation, composed of raw text from automatic text recognition (see Section 8.9) automatically aligned using the Passim tool with edited versions of the texts.²⁵ In ongoing work, we are providing initial models for the task of resolving abbreviations, notably training a ByT5 model [129] on the collected data. As well as producing models, we analyse the situation concerning the challenge of overnormalisation by the models, whereby they tend to over-standardise the outputs in addition to carrying out abbreviation resolution.

8.3.4 Low-resource and Dialectal MT

Another major challenge of MT is developing approaches to handle scenarios where there is typically little good quality data (monolingual or parallel) available, and where current LMs and MT models therefore underperform. We have developed this research direction in 2025, with several different projects tackling a range of scenarios, including translation from and into under-represented languages but which do appear

²⁵The evaluation data was manually checked and corrected to ensure its good quality.

in training data, translation from and to very low-resource languages (requiring explicit information from grammar books to carry out translation) and translation involving low-resource dialect continuums.

Translation for under-represented languages In the context of Armel Zebaze Dongmo’s PhD, supervised by Rachel Bawden and Benoit Sagot, we explored the first scenario, focusing on the use of LLMs to better translate low-resource language pairs.

We continued with the idea of few-shot prediction with the selection of semantically similar examples, a direction explored in 2024 and published in NAACL Findings 2025 [51]. In a new project, we developed the CompTra approach, published in EMNLP Findings [50], which involves decomposing each sentence into simpler phrases, translating each phrase with the help of retrieved demonstrations and finally using the original sentence and the self-generated translation pairs to produce a final translation. Our intuition was that these shorter, less complex phrases would be intrinsically easier to translate and easier to match with relevant examples. This is especially beneficial in low-resource scenarios and more generally whenever the selection pool is small or out of domain. We showed that compositional translation boosts LLM translation performance on a wide range of popular MT benchmarks across several domains.

Even when using similarity search for relevant examples, the limited amount of good quality and diverse parallel data is still a limiting factor, particularly when the language direction is into the lower-resource language. Inspired by previous work in synthetic parallel generation, we developed the TopXGen approach [52], published in EMNLP Findings. The approach consists in using an LLM to generate high quality and topic-diverse data in multiple low-resource languages, which can then be backtranslated into a high-resource source language (in our case, English) to produce useful and diverse parallel texts for in-context learning and fine-tuning of models. Our intuition was that LLMs are good at generating good quality, natural-sounding texts even in many low-resource languages and although they struggle to faithfully and consistently translate given texts into low-resource languages, they can translate well when prompted to generate into high-resource ones. We showed that the approach successfully boosts LLM translation performance during fine-tuning and in-context learning and is a promising direction for under-represented languages.

Finally, we looked at the use of reasoning in LLMs for the task of MT [90]. Since reasoning has been shown to improve LLM performance over a wide variety of tasks, but typically those involving problem-solving (mathematics and coding tasks), we sought to understand whether it could provide any gains for the MT task. It could for example help in scenarios where translation is more challenging, e.g. in low-resource scenarios and/or where the languages at play are morphologically rich and therefore could benefit from additional analysis. We did this by experimenting with several approaches involving the generation of intermediate tokens designed to help the translation process (e.g. asking the model to reason step by step, decomposing the translation as in CompTra, described above, etc.) We tested multiple language pairs and across different resource levels. We found that using intermediate “thinking tokens” does not seem to result in better MT performance. This observation was also found for the finetuning of models using synthetic chain of thought explanations; it does not perform better than standard input-output fine-tuning. However our results did point to the fact that the success of intermediate tokens during fine-tuning depends a lot on the presence of translation attempts within them. More broadly, our results suggest that using a teacher to refine target translations or to expand parallel corpora is more impactful than distilling their CoT explanations into “thinking” MT models.

Explicit learning using grammar books Different approaches must be used for translation for languages that are practically unseen in the training data of models or for which insufficient parallel data exists to train a dedicated model. In this case, there may be alternative resources, such as grammar books, that can be valuable resources for tasks such as MT. This is the case of translation from and into Franco-Provençal, our language of study for the TraLaLaM ANR project (mentioned in Section 8.8 and involving Malik Marmonier, Rachel Bawden and Benoît Sagot). Working with such a language outright is challenging, because of unstable spelling and also because of the lack of evaluation data. Therefore, in an initial step and whilst collecting appropriate data, we chose to start with a simulated setup, working with known languages that we could analyse more easily but which we encrypted to make sure that they were unintelligible to the LLMs we were working with [42]. We designed experiments with English to and from encrypted Latin and encrypted French and grammar descriptions created from Wikipedia. Contrary to previous studies, our results demonstrated that LLMs do possess a measurable capacity for explicit learning. This ability, however, diminishes as

the complexity of the linguistic phenomena involved in translation increases. Supervised fine-tuning on ad hoc chains of thought significantly enhanced LLM performance but there with little generalisation to typologically novel or more complex linguistic features.

We plan to extend the work, especially to test new LLMs as they come out and to apply the approach to Franco-Provençal when we have appropriate data. As a first step towards this goal, we produced a French version of the OLDI corpus [126], which can later be used to produce translations (and therefore parallel data) in under-represented languages and dialects of France such as Franco-Provençal [43].²⁶

MT for dialectal continuums For many lower-resourced languages and dialects, the presence of variation is an additional challenge for NLP tools. This variation can be present in the spelling system used (especially where there is not a single standard) but also in the vocabulary, morphology and syntax. For dialect continuums, the variation can be highly prevalent; NLP tools need to be robust to these high levels of variation and also capable of generating into many different varieties in a controlled and coherent way (i.e. not producing a mixture of different dialects in one prediction). This is the topic of Oriane Nedey’s PhD thesis, under the supervision of Benoît Sagot, Rachel Bawden and Thibault Clérice, with a focus on Occitan [96]. In addition to the creation of resources for Occitan (described in Section 8.8), in 2025 we began laying the foundations to test approaches for Occitan translation. Oriane Nedey produced a state of the art on MT and evaluation for language pairs involving dialect continuums and proposed a series of preliminary experiments with the Occitan continuum and current MT models [54].²⁷ The results showed a reasonable performance into English and French and a lesser performance when translating into the dialects. We also tested language identification tools, which revealed that certain models struggle to generate consistently across the continuum, especially for the lower-resource dialects.

8.3.5 Multimodal MT

A second PhD in MT was defended in 2025: Matthieu Futeral-Peter successfully defended his PhD on multimodal language modelling (with a focus on MT), supervised by Rachel Bawden, Benoît Sagot and Cordelia Schmid from the WILLOW project-team.²⁸

Two of his previously started projects were finalised and published in 2025 at international conferences: mOSCAR in ACL Findings [33] and ZeroMMT in NAACL Findings [32].

mOSCAR, also mentioned in Section 8.1.2 for aspects on multimodal language modelling, is a multimodal version of the OSCAR corpus [117], comprising web documents and associated images, which can be used to train multimodal language models. The corpus has the strength of being highly multilingual, covering 163 languages and therefore being a potentially useful source for inherently multilingual tasks such as MT. As described previously, we trained a multimodal language model on the data (in addition to more traditional caption-like resources), and one of the downstream tasks tested was MT. We showed that training on the additional interleaved data from mOSCAR²⁹ results in improved MT performance on the traditional Multi30k test set [104] (the model only trained on captioning data cannot translate at all) and is also able to effectively use images to disambiguate ambiguous texts in our CoMMuTE contrastive dataset [107].

ZeroMMT [32] is a method to train multimodal MT systems without requiring fully supervised multimodal parallel data (i.e., sentence pairs with corresponding images). The approach involves adapting a strong text-only MT model using a joint training objective: visually conditioned masked language modelling and a Kullback-Leibler divergence loss that encourages the MMT model’s outputs to remain consistent with the original MT model while integrating visual cues. This approach allows for effective learning from multimodal English data alone, making it applicable to language pairs with no existing multimodal training corpora. ZeroMMT demonstrated disambiguation performance close to state-of-the-art multimodal MT models trained on fully supervised data, as evaluated on standard MMT benchmarks and our contrastive evaluation set CoMMuTE. To further validate its generalisation capabilities, we extended CoMMuTE to three new languages: Arabic, Russian, and Chinese. Additionally, we introduced an inference-time trade-off

²⁶As also discussed in Section 8.8, we are currently in discussion with the *Institut de la Langue Savoyarde* to work together to provide corpora and tools, and for MT in particular, to produce evaluation sets.

²⁷Her paper won the best paper award at RECITAL 2025 (the French student NLP workshop).

²⁸The final version of his thesis is not yet available.

²⁹Interleaved data corresponds to documents of text with accompanying images as opposed to captioning text-image pairs where the text necessarily corresponds to the image content.

mechanism using classifier-free guidance, allowing for a controlled balance between translation fidelity and disambiguation strength without requiring additional data.

8.4 Hate Speech and Radicalisation Detection

Participants: Djamé Seddah, Chloé Clavel, Célia Nouri, Arij Riabi, Wissam Antoun, Virginie Mouilleron, Menel Mahamdi.

After the H2020 CounteR project (2021-2024), devoted to online radicalisation detection, we pursued our work on language variation through the prism of multilingual and cross-dialect hate speech. This is especially important since most multilingual language models are trained on a variety of data sources that tend to consider dialectal variations as lone instances of a given language, *de facto* ignoring important language nuances that can lead to misinterpretation of produced speech utterances or worse entailing misclassifications based on user-specific gelect varieties. Other axes we are currently investigating extend beyond the pure textual content of lone social media posts and explore various ways of exploiting their associated context (multimodal, informational, discursive, etc.).

8.4.1 Identifying Common Examples in Spanish Dialects in Hate-Speech Contexts

Our previous research on multilingual zero-shot hate speech detection [113] highlighted the cultural gap in cross-lingual settings. For instance, the Spanish word *puta* ‘prostitute’, is often used as a non-offensive intensifier but was flagged as offensive by models trained on an English dataset that focuses on misogyny. While such cultural differences across languages are increasingly studied, their impact on dialects of the same language remains unclear. After our work on hate speech contrasting different varieties of Spanish spoken either in South America or in Spain [99], we extended our approach, this time focusing on the Cuban Spanish variant and the identification of cross-dialect common examples.

As mentioned above, variations in languages across geographic regions or cultures are crucial to addressing biases in NLP systems designed for culturally sensitive tasks, such as hate speech detection or dialogue with conversational agents. In languages such as Spanish, where varieties can significantly overlap, many examples can be valid across varieties, which we refer to as common examples. Ignoring these examples may cause misclassifications, reducing model accuracy and fairness. Accounting for these common examples is therefore essential to improving the robustness and representativeness of NLP systems trained on such data. We used training dynamics to automatically detect common examples or errors in existing Spanish datasets. We demonstrated the effectiveness of using predicted label confidence for our Datamaps [125] implementation for the identification of hard-to-classify examples, especially common examples, enhancing model performance in a variety of identification tasks. Besides the original use of the Datamaps for this task, an interesting outcome lies in the introduction of a Cuban Spanish Variety Identification dataset with common example annotations developed to facilitate more accurate detection of Cuban and Caribbean Spanish varieties. To our knowledge, this is the first dataset focused on identifying the Cuban, or any other Caribbean, Spanish variety. This work was published in the VARDIAL specialised workshop, colocated with COLING 2025 [23].

8.4.2 Quantifying Geographically-Informed Sociocultural Biases

LLMs exhibit inequalities with respect to various cultural contexts. Most prominent open-weights models are trained on Global North data and show prejudicial behaviour towards other cultures. Moreover, there is a notable lack of resources to detect biases in non-English languages, especially from Latin America (Latam), a continent containing various cultures, even though they share a common cultural ground. This aspect is especially important in the context of the AeX SaLM and the ongoing collaboration with Inria Chile. As a starting point, we proposed to evaluate how culturally biased were current LLMs towards Latam. To do so we leveraged the content of Wikipedia, the structure of the Wikidata knowledge graph, and expert knowledge from social science in order to create a dataset of Question-Answer pairs, based on the different popular and social cultures of various Latin American countries. We created a database of around 20k questions and associated answers extracted from 20k Wikipedia articles, transformed into a multiple-choice questions in

Spanish and Portuguese, in turn translated to English. We use this dataset to quantify the degree of knowledge of various LLMs and find (i) a discrepancy in performances between the Latam countries, some being easier than others for the majority of the models, (ii) that the models perform better in their original language, and (iii) that Iberian Spanish culture is better known than the Latam one.

This work was carried out by Yannis Karmim and Djamel Seddah in collaboration with Luis Martí (Inria Chile) and Valentin Barriere (University of Santiago) under the Inria Chile-Inria collaboration framework and the AeX SaLM. A paper will be published as part of the proceedings of the Multilingual Multicultural Evaluation workshop collocated with EACL 2026.

8.4.3 The Counter Dataset

The first half of the Counter project faced an almost fatal challenge, namely the lack of availability of a multilingual radical content dataset covering the project’s target ideology spectrum (from white-supremacism to Jihadism). This is why we had to first develop our multitask learning architecture on adjacent domains such as hate speech before being able to focus on real radical content suitable for the law enforcement agencies that were part of the project and designed to be our end-users. Such data was made available to us by an EU-external third-party contractor, validated by the European Commission project officer, over an 18-month period during the second half of the project. The final dataset covers 12 languages, some of which were actually provided by the LEAs themselves. ALMANACH was responsible for the whole NLP work package (“Data Analytics for Detecting Radical Content”), which spanned the entire duration of the project. Due to the UE classified nature of our core deliverables (D4.3 “NLP Basic Features” and D4.4 “transfer learning for NLP”) we could not present our results on cross-lingual transfer and out-of-domain scenarios on radical content detection here. As per the consortium agreement, the only data that could be released would have to be fully GDPR-compliant, entailing a perfect pseudo-anonymisation process as demonstrated in [109]. Therefore, given the extent of the work needed for this task, we focused on a subset of languages comprising French, Arabic and English. The idea was to not only pseudo-anonymise the data-set in these three languages but also, while doing so, to keep all socio-demographic information and cultural traits that could have been linked to named-entities tied to the highly specific radicalisation domain. The goal was two-fold: (i) for a classifier trained on the new data to match the performance of a classifier trained on the original data set and (ii) to allow the redistribution of this dataset while being fully compliant to the GDPR.

It is important to note that strong anonymisation is of course necessary, but overly aggressive approaches can remove important social and contextual signals, such as usernames, locations, or references to public events, which are often crucial for detecting radical content [118]. As a result, balancing privacy protection and data usefulness is a major challenge. This is why we developed a manual pseudonymisation method tailored to our radicalisation dataset. Our approach preserves key semantic information while protecting user identities, allowing models trained on the processed data to achieve performance comparable to that of the original data. The process includes careful annotation of named entities and selective preservation of well-known public figures and events. This resulted in a multilingual, semantically rich, and privacy-compliant dataset in English, French, and Arabic, which we made freely available for research [120].

Exploring the Counter Dataset Once we properly processed the raw data provided by the contractor, we could begin to explore the challenges of detecting online radical content. As mentioned previously, the Counter dataset is a multilingual dataset designed to identify radicalisation across English, French, and Arabic texts. Unlike existing datasets that often focus on a single ideological perspective, it captures a broader range of extremist discourses, including Jihadist, far-right, and other radical discourse.

During the anonymisation phase of the dataset, we noticed a certain amount of annotation biases that could conflict with the legal context of the countries where a classifier trained on our dataset was supposed to be tested. Some of those biases were almost uniformly directed towards certain ideologies and, in some cases, towards some specific communities. Given the lack of information regarding the annotation phase by the third-party contractor (socio-demographic details of the annotators, country of origin, political stances, etc.), we decided to (i) reannotate the English and French dataset so we could contrast the original annotations with annotations coming from different people with different sociodemographic traits and political opinions, (ii) study the impact of human-label variation [119] and finally (iii) analyse our classifier results regarding different socio-demographic variables.

Our study first highlighted the complexities of human annotation in this domain, showing that annotator disagreements and socio-demographic backgrounds influence labelling decisions. By comparing prescriptive annotations (where strict guidelines are enforced) to a descriptive approach that allows for more subjective interpretations, we demonstrated how these variations impact model predictions and fairness. To further investigate the classifier biases, we generated synthetic data using a specific LLM aligned without any safeguards (Vicuña Uncensored, 13B), embedding diverse socio-demographic traits into the generated text. This persona-prompting technique enabled us to generate a 1900-document-long dataset, which we used to assess whether models treat different socio-demographic groups fairly. Our findings revealed that ethnicity, nationality, and political affiliation significantly affect classification outcomes, raising concerns about how much bias propagation can affect an NLP system such as our multitask learning classification pipeline.

Performance-wise, our results also indicate that multilingual models perform best when each language has a dedicated classifier rather than a shared one. Additionally, we challenged the definition of the task itself, contrasting classification and regression approaches, finding that classification models achieve higher accuracy but sometimes make severe misclassifications, while regression models better preserve ordinal relationships in radicalisation levels.

Beyond the technical challenges, we had to take into account major ethical concerns in the development and deployment of radicalisation detection models. Automated systems in this domain carry a risk of misuse, particularly in profiling individuals or communities based on flawed predictions. To mitigate this, in our discussion with the consortium and finally in our paper, we repeatedly emphasised the necessity of human supervision in the application of these models.

To conclude, we argue that closer collaboration between NLP researchers and social scientists is crucial for mitigating bias and improving the interpretability of radicalisation detection models. This is one of the reasons ALMANaCH is getting increasingly involved with social scientists from the Science Po's Medialab via the Salm exploratory action.

Finally, while this dataset, the first of its kind, represents an important step toward multilingual, context-aware radical content detection, it is important to note that language and behaviours associated with radicalisation continuously evolve, requiring ongoing dataset updates and model adaptations to maintain any hope of effectiveness and fairness.

This work was published at the COLING 2025 conference [47] and constitutes one of the major parts of Arij Riabi's PhD thesis [69].

8.4.4 Inferring Sociological Variation in Algorithmic Text Classification

Following our work on the CounteR Dataset, Sofia Imbert de Tremiolles has recently started her PhD, funded by the AeX SaLM and co-supervised by Djamé Seddah and Jean-Philippe Cointet (Sciences Po). A first axis of work will be to leverage existing longitudinal data from Twitter and detailed information about the socioeconomic status and political preferences of nearly 7,000 individuals that are available through an 800 million French tweets dataset (SoSweet corpus).

If robust discrepancies are observed in model quality between socioeconomic, geographical, or political backgrounds, it is possible that these variations reflect an underlying social phenomenon, or that it is simply an undesirable mistreatment of certain texts by the algorithm itself. In that case, we will develop methodology to systematically test which of these possibilities applies in each case.

In all cases, the SoSweet dataset will be used to test the effectiveness of various NLP algorithms, conditional on the authors' social, geographic, and political context. Key questions are:

- How does the reliability of toxicity detection vary across social classes?
- Is stance detection more sensitive for right-wing individuals than left-wing individuals?
- Does a cultural gap between geographic areas or different social groups make it harder to detect sexist content?

8.4.5 Abusive Language Detection in Online Conversations

In the context of Célia Nouri's PhD thesis, under the supervision of Chloé Clavel and Jean-Philippe Cointet, we contributed to the detection of abusive language in online conversations by leveraging contextual cues from

the conversation history (i.e. previous comments in the online thread) [45]. Our research demonstrated that abusive language detection models can be significantly improved by integrating conversational features that capture both the content and topology of discussions. In particular, we introduced a graph-based framework that models social media conversations as structured graphs, enabling the propagation of contextual cues across reply chains through the use of Graph Neural Networks. We compare our graph-based model with context-agnostic and flattened context-aware baselines and demonstrate that our method provides substantial improvements in abusive language detection, especially in cases where abusiveness is implicit and can only be inferred through extended interaction history. This research was funded by Chloé Clavel’s SINNet ANR project, in collaboration with the *Action Exploratoire* SaLM. The article was published at the Association for Computational Linguistics (ACL) and received the *Senior Area Chair Highlights* award, a distinction granted to approximately 4% of the accepted papers, underscoring both its methodological contribution and its impact on the field.

8.5 Fact-Checking and Disinformation Detection

Building on our prior work in hate speech detection and radicalisation analysis, where understanding implicit narratives, entity relations, and ideological framing is essential, we extended these insights to the task of automated fact-checking. Similar to radical content, misinformation often relies on subtle semantic cues and complex relationships between actors, events, and claims that are difficult to capture with surface-level text features alone. In this research direction, we present a cross-modal approach that infuses language models with semantic graph representations to address these challenges. By combining contextual linguistic information with structured semantic knowledge derived from Abstract Meaning Representation (AMR) graphs and Wikidata, we capture deeper relationships between entities and events. Our experiments on a large multilingual dataset in English and German show that this integrated approach consistently outperforms unimodal baselines and generalises better to out-of-domain data.

Our method represents each claim using both its textual form and an AMR-based semantic graph. We enhance AMR graphs by mapping PropBank roles to VerbNet and linking named entities to Wikidata in order to incorporate external knowledge. Graph representations are encoded using Graph Attention Networks, while textual inputs are modeled with pretrained transformers such as RoBERTa. We then combine these modalities using the GreaseLM framework, which enables deep, bidirectional interaction between language and graph representations across multiple layers. This fusion mechanism allows structured knowledge and contextual semantics to mutually reinforce each other. We evaluate our approach against strong transformer and GNN baselines, and demonstrate through ablation and out-of-domain experiments that deep cross-modal fusion leads to improved robustness and performance. This work marks a milestone in our ongoing research on language modelling, as it shows that using structured semantic knowledge (the AMR graphs) jointly with discriminative LLMs improves the performance of complex downstream tasks such as fact checking. This work, published at the specialised workshop Multimedia AI against Disinformation (MAD’25) colocated with ICMR’25 [36], was conducted as part of an ongoing collaboration with Zehra Melce Hüsünbeyi, her PhD advisor, Tatjana Sheffler from Ruhr-Universität Bochum, and Djamé Seddah. Melce spent 6 months at ALMANACH as part of a 6-month PhD “cesure” as a research engineer, funded by the BPI Scribe Project. A multimodal fact-checking corpus with human annotation and LLM judgments has also been developed, and the paper describing the work is currently under review.

8.6 Dialogue Modelling

Participants: Chloé Clavel, Ha Anh Ngo, Nicolas Rollet, Aina Garí Soler, Chenwei Wan.

8.6.1 Repair Modelling

Regarding the scope of Anh Ha Ngo’s PhD, funded by the Paris Île-de-France Région through the DIM AI4IDF and under the supervision of Chloé Clavel, Catherine Pelachaud (Sorbonne Université, ISIR) and Nicolas Rollet (Telecom-Paris, on secondment in ALMANACH), our research focuses on modelling

conversational repair mechanisms to improve human-agent interaction. Building upon our initial linguistic analysis in the first year, [114] we published a novel multimodal approach for detecting human repair initiation in the EMNLP 2025 paper, “Mm, Wat?” Detecting Other-initiated Repair Requests in Dialogue [44]. This work proposed a computational model that integrates retrained text and audio embeddings with handcrafted linguistic and prosodic features grounded in Conversation Analysis literature. Specifically, the results suggested that prosodic cues, such as pausing behaviours and voice intensity, synergise with linguistic patterns, such as coreference usage and repetition, play an essential role in identifying when a user signals trouble and initiates a need for repair.

To expand this research by evaluating the generalisability of repair detection across different languages and models, in the second half of 2025, we submitted a study to LREC 2026 involving the annotation of other-initiated repair in the French NOXI corpus. We assessed the capabilities of LLMs to serve as both automated annotators and explainers for the disagreements in annotation between LLMs, human annotator and human experts.

8.6.2 Word Meaning Negotiation

As part of Aina Garí Soler’s postdoctoral research, and towards the end of the year in her capacity as an associated member, we continued our investigations on word usage dynamics in dialogue. This work was supervised by Chloé Clavel and Matthieu Labeau (Telecom-Paris) and funded by Chloé Clavel’s ANR PRME SINNet project. Our submission to the LRE journal (Language Resources and Evaluation) on the NeWMe corpus is now available as a preprint [89] and the corpus is available online. The NeWMe (Negotiating Word Meaning) corpus is a collection of annotated Word Meaning Negotiation (WMN) sequences. WMNs are parts of conversations where a speaker raises a clarification request or an objection about the usage of a specific word, prompting a metalinguistic discussion aimed at clarifying or refining the word’s meaning. The corpus contains various kinds of interactions (written, oral, dyadic, and multi-party) and is annotated with WMN sequences, their components, and related phenomena.

NeWMe is the first corpus of its kind, but it is of limited size, and we have worked on developing methods for its semi-automatic extension. WMNs have three key components: the trigger, which is a problematic word usage; the indicator, an utterance signaling a problem with a word’s meaning; and the negotiation—one or more conversational turns where participants address the misunderstanding or disagreement. We have specifically explored the automatic detection of WMN indicators in conversation. Our proposed models have a much better precision than previous regex-based identification methods, and show promise for accelerating the WMN identification process, but there is still room for improvement and we identify relevant directions for future work. This work resulted in an article accepted at Findings of EMNLP 2025 [35].

To better understand the mechanisms by which a word can be misunderstood or disputed, we paid particular attention to the trigger component. We conducted an extensive literature review, identifying linguistic factors that contribute to potentially problematic word usages, as well as computational methods and data that facilitate their detection. This review connects works from various disciplines and highlights areas for future research. The outcome of this work is a survey that has been accepted to *SEM 2025 [34].

Finally, we focused on scare quoted word usages as examples of potentially problematic word usages that could trigger a WMN. Quotation marks have a wide range of uses in written language. One of them is scare quoting: scare quotes mark irony, distance, or disagreement about word meaning or lexical choices. As such, they are explicit markers of the writer’s acknowledgment that a word may be problematic and their anticipation of a possible misalignment in meaning. We created the first annotated corpus of quoted word and phrase usages focused on the scare vs non-scare quotes distinction, and in their use in conversation. This work is currently under review at LREC 2026 and it will contribute to the detection and characterisation of problematic word usages.

8.7 Natural Language Processing for Specialised Domains

Participants: Chloé Clavel, Benoît Sagot, Éric Villemonte De La Clergerie, Lucie Chenain, Rian Touchent, Simon Meoni, You Zuo.

8.7.1 Biomedical and Clinical Domains

During her PhD, Lucie Chenain supervised by Chloé Clavel and Anne-Catherine Bachoud-Levi (APHP, ENS, Inserm), introduced emoHD, the first speech corpus designed to study emotional expression and psychiatric symptoms in Huntington’s Disease (HD) [15]. Addressing a major clinical challenge in HD, this work investigated how emotions expressed in speech relate to the severity of psychiatric symptoms, including depression, irritability/aggressivity, apathy, and obsessive/compulsive symptoms. HD participants first underwent a clinical assessment in which neurologists evaluated these psychiatric symptoms using standardised measures. They then completed a recorded speech interview conducted by a neuropsychologist. These recordings were then emotionally annotated using a three-level annotation scheme, moving from general emotional descriptors (emotional activation), to basic (primary emotions), and finally to fine-grained affective states (affective phenomena). The analyses reveal a global reduction of emotional expressiveness in HD compared to healthy controls. Expressed emotions aligned with patients’ psychiatric profiles (e.g. anxious and nervous speech with depression; frustrated speech with irritability and aggressivity), highlighting speech as a promising marker for emotional and psychiatric symptoms, paving the way for remote monitoring and personalised care in HD. This PhD work is the result of a collaboration between Inria, the Ecole Normale Supérieure, and Inserm, and is funded by Université Paris Cité (ED 474).

Through several initiatives (BPI-funded projects ONCOLAB and FG4H, collaboration with AP-HP, and the INRIA AOC action), we continued to investigate the use of NLP techniques, particularly LLMs, applied to medical documents. This domain presents specific challenges: clinical data (Electronic Health Records, EHRs) are highly sensitive, difficult to access, and especially scarce in French. Conversely, medical knowledge is more readily available through ontologies, scientific publications, and guidelines. To address these constraints, we are progressively developing a new methodological paradigm based on:

1. the generation of realistic but non-confidential fictitious clinical data;
2. the automatic annotation of this synthetic data using LLMs;
3. the training of Small Language Models (SLMs) that can be securely deployed on private infrastructures.

As part of this effort, Rian Touchent finalised Biomed-Enriched, a two-step pipeline for paragraph-level annotation of large public biomedical corpora such as PubMed and ISTEEX. In the first step, Llama-3.1-70B annotates a subset of PubMed paragraphs by type, domain, and educational quality. These annotations are then used to train an SLM (XML-RoBERTa), which subsequently annotates the full PubMed corpus. This enriched annotation enables fine-grained selection of pre-training datasets, for example targeting high-quality clinical case descriptions. Experiments demonstrate that an encoder-based SLM trained on these curated subsets achieves performance comparable to BioClinical-ModernBERT on 11 clinical benchmarks, while requiring 2.5 times less pre-training data, highlighting the value of informed data selection.

Data selection techniques were also used by Rian Touchent to train a new version of CamemBERT-bio (v2) [127, 128] and to develop ModernCamemBERT-bio, a French biomedical encoder based on ModernBERT/ModernCamemBERT with an extended context window (up to 8,192 tokens). Such extended context capacity is particularly relevant for processing long clinical documents or full patient histories. Optimised GLiNER variants of these models are also under development for information extraction tasks (ONCOLAB context) and ICD-10 coding (collaboration with AP-HP through the co-supervision of Anh Thu’s M2 internship).

For experiments on ICD-10 coding, models were trained on several thousand synthetic hospital reports generated by an LLM (Mistral) using clinical scenarios and detailed medical guidelines. In parallel, Rian Touchent explored the reformulation of publicly available PubMed clinical cases into standard hospital report formats. The original case provides the narrative structure and essential medical content, which the model then adapts into the style of clinical notes.

Another line of research concerns the generation of synthetic facsimile clinical documents, conducted as part of Simon Méoni’s PhD work. The approach consists in extracting non-confidential UMLS concepts from real documents and using them as inputs to train a generator model. Reinforcement learning is employed to optimise the similarity between facsimile and real documents, thereby improving realism without compromising confidentiality. A complete workflow is now available for producing such facsimile documents, particularly on the English MIMIC dataset, and was further validated through the co-supervision of Tripodi’s AP-HP internship. The objectives included generating French facsimiles and using them to train multi-label

ICD-10 classifiers. Due to French regulatory constraints, the methodology prioritised synthetic documents authored by clinicians rather than real patient data, thereby minimising re-identification risks. A significant part of Simon Méoni’s work in 2025 was dedicated to evaluating the privacy risks associated with facsimile generation. Although facsimiles contain no explicit identifiers, re-identification could still arise from unique sequences of UMLS concepts. Several privacy metrics were identified and tested, leading to the development of mitigation procedures designed to reduce potential confidentiality leaks while preserving the utility of the generated data for downstream tasks.

8.7.2 Patents

Our research on NLP for patents, carried out in collaboration with Questel (formerly with qatent, acquired by Questel in 2024) in the context of You Zuo’s CIFRE PhD thesis (industrial supervisor: Kim Gerdes), focuses on developing representation learning methods that respect the unique structure, length, and technical density of patent documents, with a particular emphasis on scalable and annotation-light retrieval. In 2025, our first line of work focused how self-supervised contrastive learning can be adapted to the patent domain. We showed that standard SimCSE-style dropout augmentation leads to overly uniform embeddings for patents, due to their length and internal redundancy. To address this, we introduced section-based contrastive learning, where different patent sections (e.g. abstract, claims, background) are treated as complementary semantic views. This approach exploits patents’ intrinsic discourse structure to learn embeddings that preserve both global semantic coherence and local continuity, achieving state-of-the-art performance in prior-art retrieval and classification without relying on citation or IPC (International Patent Classification) supervision [91].

Complementing dense representations, our second line of work explores geometry-aware sparse representations for patent retrieval. We initiated the development of an unsupervised discretisation of dense embedding spaces using overlapping hyperspherical coverings, producing learned sparse vectors that capture both exact terminology and semantic neighborhoods. These representations are directly compatible with inverted indexes and significantly outperform traditional lexical and clustering-based sparse baselines, while combining effectively with dense retrievers. Together, these contributions advance robust, interpretable, and scalable retrieval methods tailored to complex patent data. We intend to publish on this research in 2026.

8.8 Corpus and Tools for Languages of France

Participants: Benoît Sagot, Lucence Ing, Thibault Clérice, Rachel Bawden, Rasul Jasir Dent, Juliette Janès, Oriane Nédey, Malik Marmonier.

The Inria-funded COLaF project (2023–2027) is dedicated to the collection and development of NLP tools and resources for French and the other languages of France, with particular attention to less-resourced languages and varieties [55]. The project covers textual, audio and video data, with the aim of providing corpora and tools for written, spoken and signed language. It includes the collection, normalisation and documentation of pre-existing data, including data that are currently inaccessible or unusable for research purposes, as well as the development of NLP tools tailored to these languages, such as tools for linguistic annotation and MT. ALMAnaCH is in charge of the textual modality and the MULTISPEECH team (Inria Nancy) of the speech and sign language modalities. The two PIs of the project are Benoît Sagot and Slim Ouni (MULTISPEECH). Lucence Ing’s activities includes that of COLaF project manager for ALMAnaCH.

In 2025, following our previous work, we carried out work both at an abstract level (explicit learning for low-resource languages; MT-related aspects are covered in Section 8.3 dedicated to MT), and at a more concrete level, covering data encoding issues as well as language resource development for a number of languages of France.

We also continued and, sometimes, initiated, discussions and collaborations with multiple organisations dedicated to languages of France other than French, in particular the Agence Régionale de la Langue Picarde (Picard), Lo Congrès (Occitan), l’Institut de la Langue Savoyarde (Savoyard being a variety of Arpitan/Franco-Provençal) and the Bibliothèque Numérique de la Sorbonne (NuBIS). The work on COLaF is also carried out in close collaboration with the ANR project Tralalam (also discussed in Section 8.3). Persée, one of the collaborating institutions, provided feedback regarding the introduction of our LADaS work into their editorial pipeline to publish academic archives, with an 85% productivity increase over three months for

the specific task of document layout annotation (See Section 8.9.2 and 8.8.6 for more details on LADaS in 2025).

An important event for the project in 2025 was the yearly COLaF meeting, where around 50 people, members of the project and representatives of partner institutions met at Inria Paris in June for a one-day-long workshop.

8.8.1 French OLDI

In collaboration with the Tralalam project, Malik Marmonier, supervised by Rachel Bawden and Benoît Sagot, developed the first French partition of the OLDI Seed Corpus [111], our submission to the WMT 2025 Open Language Data Initiative (OLDI) shared task [43]. Its creation process involved using multiple MT systems and a custom-built interface for post-editing by qualified native speakers. The source data presented unique translation challenges, as it combines highly technical, encyclopedic terminology with the stylistic irregularities characteristic of user-generated content taken from Wikipedia. This French corpus is not an end in itself, but is intended as a crucial pivot resource to facilitate the collection of parallel corpora for the under-resourced regional languages of France, as discussed previously in Section 8.3.4.

8.8.2 Occitan Dialects

Our work on Occitan corpus development, in the context of Oriane Nédey’s PhD thesis supervised by Benoît Sagot, Rachel Bawden and Thibault Clérice, addresses the dual challenge of data scarcity and fine-grained linguistic variation across the Occitan dialect continuum. In 2025, a first line of research focused on building ethically sound, linguistically rich resources from UGC. We designed and documented a full pipeline for data collection, cleaning, and anonymisation of forum data, ensuring compliance with ethical standards while preserving crucial sociolinguistic information [66]. This methodology underpins ForumOccitania, a large-scale corpus of online discussions that combines textual data with self-declared metadata on dialect, geography, age, and speaker profile [84]. Beyond corpus construction, we carried out quantitative, qualitative and exploratory computational analyses, showing that the corpus is predominantly written in Occitan using classical orthography, largely by young neo-speakers, and that it exhibits clear dialectal signals from major varieties such as Lemosin, Lengadocian, Gascon, and Provençau.

Complementing forum data, we developed OcWikiDialects, a large encyclopedic corpus drawn from Occitan Wikipedia, enriched with explicit dialect labels covering eight major varieties and two transitional ones. With millions of tokens and multiple textual representations and segmentations, this resource broadens the empirical basis for studying written Occitan variation. Together, these corpora provide a coherent, methodologically grounded infrastructure for sociolinguistic and NLP research on Occitan and its varieties.

8.8.3 Picard

In 2025, we continued the collaborations established with the Agence régionale de la Langue Picarde (ARLP) and the Université de Picardie Jules Verne (UPJV). UPJV provided the content of its textual database of documents in Picard, **PICARTEXT**, to COLaF. It constitutes the largest dataset currently available for the language, spanning from the 18th to the 21st centuries and comprising 6 million tokens. Work is underway to transform this plain-text database into the COLaF XML-TEI format, making it structured and searchable. At the same time, UPJV is annotating the dataset with linguistic information in order to train language annotation models (e.g. lemmatisers), using the **Pyrrha application** (a tool for post-correction of lemmatised and morphosyntactic tagged corpora) on the instance provided by COLaF on Huma-Num server.

In parallel, a digitisation campaign organised by Juliette Janès was carried out at ARLP to recover analogue documents from their archives, namely the Picard literature competitions from 2005 to 2020 (approximately 40 documents per year). These materials are currently being processed in XML-TEI using the LADaS and ATR pipeline under development.

8.8.4 French-Based Creoles

Our contributions to French-based Creole studies combine computational linguistics with historical linguistics to reassess the development of French-based Creole languages through large-scale textual evidence. This work is carried out primarily by Rasul Jasir Dent in the context of his PhD, supervised by Benoît Sagot,

Thibault Clérice and Pedro Ortiz (Common Crawl). A key challenge in this regard is the creation of new digital corpora designed to capture contact-induced variation across time and space. In 2025, building on the Molyé corpus, which brings together early Creole attestations and European contact varieties over four centuries [100], we have worked on corpus construction methods tailored to sparsely documented languages. By reframing language identification as a data-mining task rather than full multi-class classification, we were able to rapidly assemble broad and representative corpora for several French-based Creoles, prioritising coverage over exhaustive document processing [74].

These resources underpin our historical investigations into the relationships among Atlantic and Indian Ocean Creoles. Based on our corpora, we provide evidence for a nineteenth-century transoceanic continuum ranging from normative French to basilectal Creole, with intermediate “Para-Creole” varieties showing feature-level diffusion across regions [62]. This corpus-driven perspective also informs our re-evaluation of French-based pidgins. A comparative analysis of Français Tirailleur and Tây Bôï, supported by newly analysed textual data, suggests that both were shaped by top-down pedagogical practices rather than emerging solely from spontaneous communicative necessity [61].

Finally, our corpus-building efforts have extended to contemporary NLP infrastructure through shared evaluation campaigns, resulting in new public datasets for multiple Creole languages and fostering collaboration around their digital documentation [48].

8.8.5 Language Identification

In 2025, again in the context of Rasul Jasir Dent’s PhD work, we redefined the language identification task from the perspective of corpus creation for less commonly written languages: rather than treating language identification as a standard multi-class classification problem, we argued that it is more productively framed as a targeted data mining task, in which only a small fraction of available documents are expected to be relevant [74, 31]. Building on this reframing, we designed an efficient pipeline that prioritises the rapid exclusion of irrelevant material, thereby minimising computational and human resources spent on uninformative data, while making it possible to find data for very scarcely attested language varieties (a “needle-in-a-haystack” scenario). Our approach makes it possible to filter massive collections within a few hours, supporting the scalable construction of new digital corpora [31]. Although we primarily applied the methodology to French-based Creole languages (see above), our proposed framework is generic and applicable to language identification scenarios characterised by extreme class imbalance, and we carried out preliminary experiments to minority regional languages of France.

8.8.6 Data Encoding

We resumed our joint work on methods for encoding and representing linguistic diversity in text corpora, especially for under-resourced, non-standard and historically situated varieties. In 2025 we developed a rigorous TEI-based schema that captures fine-grained metadata for language variants across temporal, geographic and sociolinguistic dimensions, addressing a notable gap in existing standards such as BCP47 by incorporating richer identifiers via resources like Glottolog [37]. This schema, whose main developer is Juliette Janès, plays a central role in COLaF, as our aim is to apply it to all corpora distributed in this context.

Complementing this structural work, we explored the diachronic and contact-influenced nature of language variation, exemplified by work on French-based Creole (see above), which uses digitally compiled textual evidence to reconceptualise these varieties within networked historical contexts rather than strict genealogical models [64].

8.8.7 The Parallel Corpus of the Parable of the Prodigal Son

In order to cover the diversity of the regional language varieties spoken in metropolitan France, in 2025 we started to build a parallel corpus of the Parable of the Prodigal Son, a collection of translations of this parable produced for national linguistic surveys in the 19th century. We collected the data with the help of two external researchers, Sven Ködel and Alexandre Génadot, structured the data and language metadata in XML-TEI files (See Section 8.8.6), aligned the versions and established a comparison between these data and the *Atlas Linguistique de France* ones by projecting lexical data on georeferenced maps. The work is still in progress (increase of data, improvement of the automatised full pipeline) and a publication is upcoming.

8.9 Automatic Text Recognition for Historical Documents

Participants: Thibault Clérice, Benoît Sagot, Hugo Scheithauer, Alix Chague, Juliette Janès, Lucence Ing, Benjamin Kiessling, Hassen Aguli, Nicolas Angleraud, Antonia Karamolegkou.

Research on Automatic Text Recognition (ATR) for historical documents continued actively in 2025 and can be structured around four complementary axes: (1) a critical and epistemological reflection on ATR and transcription practices, (2) the consolidation and incremental improvement of existing models, tools, and standards, (3) the dissemination and large-scale application of previously developed resources, in particular transcription datasets, and (4) the exploration of new and challenging directions for text recognition and document structuration.

8.9.1 Epistemic approach to ATR

The year 2025 was marked by a strong epistemological engagement with the meaning, scope, and implications of transcription and automatic transcription in the humanities. This work resulted in several publications, including the preparation of a collective book and multiple research articles.

The book *Apprendre à lire aux machines* [81], currently available as a preprint and co-edited by Thibault Clérice, Alix Chagué, and Ariane Pinche (CNRS), constitutes the first French-language volume devoted entirely to ATR for the humanities. It brings together contributions on seven writing systems (Latin, Greek, Armenian, Arabic, Chinese, Bengali, and Urdu), alongside four transversal chapters addressing broader issues such as ATR infrastructures, methodological choices, and the historical development of the field. This publication was complemented by further studies examining the impact of ATR on humanities research [16] as well as its current limitations in relation to model architectures and available datasets [20, 65].

The year concluded with a preprint by Benjamin Kiessling offering a fine-grained analysis of transcription guideline variability in the humanities, the tension between traditional scholarly transcription and transcription optimised for ATR, and the implications of these choices for model generalisation across scripts [80].

8.9.2 Improvement and Consolidation of Existing Results

In contrast to previous years focused on rapid methodological advances, 2025 was primarily dedicated to the stabilisation, consolidation, and refinement of existing tools, models, and standards. This included the publication of a reference paper on Kraken V5 [38], the ATR engine to whose development Inria has significantly contributed, as well as improvements in dataset documentation through enhanced schemas for HTR-United [21] and a revised version of the LADaS annotation guidelines [79].

Additional work addressed the transcription of modern scripts in a European context over the last five centuries [63]. For medieval material, data production continued within the HTRogène project, although without new publications, apart from the release of updated models (CATMuS Medieval 1.6.0, [97]). Finally, ongoing but as yet unpublished work focused on improving layout analysis within Kraken and eScriptorium; tangible results from these efforts are expected in 2026.

8.9.3 New Challenges in ATR

Beyond consolidation, four emerging challenges were explored in 2025, although they have not yet resulted in substantial published output.

The first challenge arose within the *Back In Time* project and concerns the automatic transcription of enciphered texts composed of unique or highly idiosyncratic glyphs. Addressing this problem required pushing ATR methodologies beyond standard assumptions about scripts and Unicode-available signs. This work led to the development of *Glyphea*, an extension for eScriptorium specifically designed to handle complex, non-standard scripts and enciphered writing systems, as well as to broader improvements to eScriptorium in this direction.

The second challenge concerns the development of more generalisable and language-independent ATR models. In this context, ALAMAnCH is organising the ICDAR 2026 Competition on Multilingual Medieval

Handwriting Recognition, coordinated by Benjamin Kiessling and Thibault Clérice, with the goal of fostering methodological advances in cross-lingual and cross-script generalisation.

The third challenge focuses on the automatic transcription and structural formalisation of edited Ancient Greek texts. Conducted within the Corpus Liberatum Linguae Graecae project (Thibault Clérice, Nicolas Angleraud, Antonia Karamolegkou, Benoît Sagot), this work evaluates the ability of both ATR systems and Visual LLMs (VLLMs) to recover structural information and Ancient Greek text from printed scholarly editions [57]. In 2025, foundational work was completed through the creation of a synthetic dataset, with a publication planned for early 2026. A fourth challenge is being addressed within the PhD project of Hugo Scheithauer. His research focuses on Optical Music Recognition (OMR) for historical documents and will begin to yield published results in 2026. This work also includes the release of a first dataset dedicated to the transcription of 19th-century musical manuscripts. The PhD is conducted under the supervision of Thibault Clérice, Gonzalo Romero-García (Epita), and Laurent Romary.

8.9.4 Application of ATR at scale

Finally, 2025 represented a turning point in the large-scale application of ATR expertise through the creation of CoMMA (Corpus of Multilingual Medieval Archives). This corpus comprises approximately 33,000 manuscripts automatically transcribed using CATMuS models over a four-month period, in collaboration with EquipEx Biblissima+ and three humanities researchers [73].

The resulting corpus, amounting to roughly 3.3 billion tokens of Latin and Old French texts spanning the 9th to the 16th centuries, is accessible through a dedicated user interface, *Textile* (see the [CoMMA web site](#) for more details). *Textile* relies on the Distributed Text Services protocol [59, 60] to enable scalable access and exploration. Work is currently underway to extend CoMMA to additional digitisation repositories and to other languages, notably from the Iberian Peninsula, in 2026.

8.10 NLP for Historical and Literary Sources

Participants: Thibault Clérice, Lucence Ing.

Our work on earlier stages of languages and, more broadly, on NLP for the computational humanities continued in 2025. This activity combined a sustained interest in stylometry with research on morphosyntactic tagging for historical language stages, as well as a growing set of investigations in computational philology.

Regarding stylometry, two papers were published in the context of the Computational Humanities Research Conference (CHR 2025), a leading venue in the field, held in Luxembourg. The first paper presented a methodological study on the impact of temporal distance and stylistic change in an author's literary production [28], based on a carefully controlled corpus. Since stylometric methods are frequently applied to works with uncertain or entirely unknown dates, this analysis of the effect of time gaps between texts contributes to a critical reassessment of recall and robustness in stylometric attribution.

The second contribution focused on a previously anonymous text, the *Life of Saint Lambert*, which in earlier work by Thibault Clérice, Ariane Pinche and Jean-Baptiste Camps had been associated with Wauchier de Denain in unsupervised settings, as an unexpected outcome of a larger study of Wauchier's known works. This new and more comprehensive study [29], based on evidence drawn from ten manuscripts, demonstrates a strong probability in favour of this attribution. Beyond the specific case study, the paper proposes a methodological framework for addressing authorship questions in large, predominantly anonymous medieval manuscript corpora.

Computational medieval philology has emerged as a recent extension of the team's research activities, notably following the arrival of Lucence Ing and the progressive articulation of ATR outputs with downstream philological analysis (such as the study of Saint Lambert's life). Rather than limiting ATR to transcription alone, this line of work aims to open new research avenues by making automatically produced texts usable for philological inquiry. One of the first developments in this direction concerns the normalisation of ATR outputs within the context of the ParamHTRs project (see Section 8.3). In the same area of manuscript data post-processing, Lucence Ing worked with two external researchers, Matthias Gille Levenson and Carolina Macedo on the automatic multilingual segmentation of raw ATR outputs in order to enable the alignment of

texts across languages, with a particular focus on the alignment of textual variants [95]. This line of research aims to support a broader and more systematic approach to the study of textual transmission within medieval traditions.

Finally, Lucence Ing and Thibault Clérice further developed our work on morphosyntactic annotation through their collaboration within the binational project E-CaM (“Étiquetage lexico-grammatical du castillan médiéval”). This project focuses on the design of an open dataset and corresponding guidelines tailored to historical variants of Castilian, with the aim of fostering reusable and interoperable resources for diachronic linguistic analysis [75].

8.11 Multimodal Approaches to Human-agent and Human-human Interaction

Participants: Justine Cassell, Sinem Demirkan, Hasan Onur Keles, Cindy Evellyn De Araujo Silva, Reem Al Najjar, Sophie Etling, Gabrielle Alimi, Barokshana Baskaran, Zofia Milczarek, Biswesh Mohapatra, Marius Le Chapelier, Clara Coridon, Marie Nsingi Kinkela, Yassine Machta, Mayank Palan, Oussama Silem, Erinda Morina, Rémy Ben Messaoud, Justine Reverdy.

This research direction advances understanding of human-like communication by studying how interpersonal dynamics—like synchrony, rapport, and personality expression—affect collaboration (in particular transactivity) and interaction, both in humans (including children) and between humans and AI-driven conversational agents

8.11.1 Using Interbrain Synchrony and Rapport Building to Understand Productive Peer Collaboration

As part of our large-scale research programme on collaboration, our interdisciplinary team continued and extended its investigation of dyadic collaboration in children aged 5 to 12. The overarching objective of this work is to better understand the mechanisms underlying successful collaboration by examining the relationships between interpersonal rapport, collaborative performance, and inter-brain synchrony (IBS), measured using functional near-infrared spectroscopy (fNIRS) hyperscanning. Building on the experimental framework established previously, the study focuses on how children’s language and non-verbal behaviours relate to both learning outcomes and neural coupling between interacting peers during collaborative problem-solving tasks. In 2025, the project introduced a focus on transactivity, a key dimension of collaboration defined as the extent to which participants take up, respond to, and build upon one another’s ideas during interaction. Within this framework, the team conducted a longitudinal, multimodal data collection in Parisian primary schools. Data were collected from 13 dyads of children over a four-week period using fNIRS hyperscanning. In addition, 8 dyads from another age group were recorded using the same protocol, enabling replication and comparison across three distinct age groups. This expanded dataset enables us to advance the preliminary investigation of the relationship between transactivity, learning gains, and inter-brain synchrony.

8.11.2 Son of Sara: Developing a new LLM-based Embodied Conversational Agent

As part of our ongoing work on socially capable conversational agents, we continued the development of Son of SARA, an embodied conversational agent designed to support natural and effective interaction with human users. The project aims to equip dialogue agents with both verbal and non-verbal interactional skills that are essential for collaborative communication. In 2025, the work focused on extending the agent beyond speech-only interaction by adding a visual and embodied dimension to the conversation. Building on an existing conversational framework enabling spoken dialogue, we developed the agent’s virtual body and implemented the software infrastructure required to synchronise non-verbal behaviours with speech. This included the design and implementation of preliminary rule-based models for the generation of facial expressions and gestures, allowing the agent to produce visible communicative signals aligned with its speech turns. In parallel, the project advanced the agent’s conversational dynamics by improving its turn-taking capabilities. A predictive model based on voice activity was integrated to enable the agent to anticipate turn transitions during interaction. This contributes to more interactive and rhythmically natural

exchanges, bringing agent-human conversations closer to the temporal structure of human-human dialogue. Together, these developments mark a significant step toward the fully embodied nature of the Son of SARA conversational agent. They provide a solid foundation for future work on data-driven, real-time gesture generation conditioned on vocal features, as well as on further refinements of turn-taking models, including adaptation to additional languages.

8.11.3 Conversational Grounding in Dialogue Systems

As part of our research on socially competent dialogue systems, we continued our investigation of conversational grounding, a fundamental mechanism through which interlocutors establish and maintain mutual understanding over the course of an interaction. This work focuses on the ability of LLMs and LLM-based agents to represent, update, and exploit shared knowledge during dialogue. In 2025, the research shifted from grounding limited to immediate conversational context toward the modelling of long-term common ground, that is, information accumulated and reused across extended interactions. To support this objective, we introduced Indiref, a new benchmark designed to assess the use of grounded information. This gave us a way to test different representation techniques for common ground including textual and visual representations. Our evaluation revealed that standard representation techniques performed poorly on this benchmark. In response, we developed a pipeline to generate synthetic situated conversations, which were then used to train LLMs via reinforcement learning. We successfully demonstrated that this approach significantly enhances grounding performance, allowing us to isolate the key requirements for establishing robust long-term capabilities in conversational agents. All of this work is described in detail in an article to appear in 2026; a preprint is already available [83].

8.11.4 Exploring Interpersonality: Multimodal Personality Cues in Embodied Conversational Agents

As part of our research on personality expression in embodied conversational agents, we continued our investigation of interpersonalit, a framework that captures how the multimodal manifestation of personality emerges from interaction and is shaped by the interlocutor, the communicative context, and broader situational factors. In 2025, we successfully obtained ethical approval from COERLE, enabling the launch of a new experimental study involving the collection of a corpus of quadrads (groups of four participants). The experimental design was substantially revised in close collaboration with the KETI team, with the objective of enabling optimal observation of gestures and embodied behaviors, and ensuring the creation of a replicable corpus. This design choice lays the groundwork for future cross-cultural analyses of interpersonalit and multimodal interaction dynamics. This work establishes the methodological and empirical foundations for future analyses of interpersonalit and multimodal interaction dynamics, and supports the long-term objective of understanding how personality is jointly constructed through social interaction across cultural contexts.

9 Bilateral contracts and grants with industry

opensquare

Participants: Benoît Sagot.

Partner type: Inria start-up

Leader for ALMAnaCH: Benoît Sagot.

Dates: 1 Dec 2016–present

Description: Opensquare was co-created in December 2016 by Benoît Sagot with 2 senior specialists of human resources (HR) consulting. It is dedicated to designing, carrying out and analysing employee surveys as well as HR consulting based on these results. It uses a new employee survey analysis tool, enqi, which is still under development. This tool being co-owned by opensquare and Inria, both parties have signed a Software Licence Agreement in exchange for a yearly fee paid by opensquare to

Inria based on its turnover. Benoît Sagot currently contributes to opensquare, under the “Concours scientifique” scheme.

META

Participants: Benoît Sagot, Djamé Seddah, Pierre Chambon, Romain Froger.

Partner type: Company

Leader for ALMANACH: Benoît Sagot.

Dates: 1 Jan 2018–present

Funding received: €331,260

Description: Our collaboration with META AI is centered around the joint supervision of CIFRE PhD theses. A first collaboration (Louis Martin’s PhD thesis), co-supervised by Benoît Sagot, Éric de La Clergerie and Antoine Bordes (META) was dedicated to text simplification (“français Facile À Lire et à Comprendre”, FALC), in collaboration with UNAPEI. This collaboration was part of a larger initiative called Cap’FALC involving (at least) these three partners as well as the relevant ministries. Louis defended his PhD in 2022. Two other joint PhD theses started in 2021 and were defended in 2024. Paul-Ambroise Duquenne’s PhD, co-supervised by Benoît Sagot and Holger Schwenk (META), is dedicated to sentence embeddings for massively multilingual speech and text processing. Tú Anh Nguyen’s PhD, co-supervised by Benoît Sagot and Emmanuel Dupoux (META), is dedicated to the unsupervised learning of linguistic representations from speech data, with a focus on textless dialogue modelling and speech generation. Pierre Chambon’s PhD is dedicated to code generation with language models. Finally, Romain Froger, supervised by Djamé Seddah and Thomas Scaliom, started his PhD last Fall, working on reinforcement learning in multi-agent scenarios.

In addition, Benoît Sagot received in 2024 a \$50,000 gift grant from META AI in the context of the release of the LLAMA 3 series of models.

Qatent (now part of Questel)

Participants: You Zuo, Benoît Sagot, Éric de La Clergerie.

Partner type: Former Inria start-up, now part of a company

Leader for ALMANACH: Benoît Sagot.

Dates: 1 Jan 2021–present

Description: Qatent is a former startup supported by the Inria Startup Studio and ALMANACH that applies NLP to help write better patents faster. Its creation followed the 18-month secondment (“détachement”) at ALMANACH of Kim Gerdes, one of the three founders of the company, and benefitted from ALMANACH’s scientific expertise and the Inria Startup Studio’s counselling and financial support. It also led to You Zuo’s CIFRE PhD thesis, co-supervised by Benoît Sagot, Éric Villemonte De La Clergerie and Kim Gerdes (now at qatent), which continues the collaboration. In 2024, qatent was acquired by Questel, a global leader in intellectual property (IP) management and technology services, without a significant impact on You Zuo’s PhD.

AFNOR

Participants: Chloé Clavel, Xiangyu An.

Partner type: Company

Leader for ALMAnaCH: Chloé Clavel.

Dates: 2 Jun 2025–2 Jun 2028

Funding received: €70,000

Description: The collaboration between ALMAnaCH and AFNOR focuses on a joint CIFRE thesis dedicated to the evaluation of conversational systems.

9.1 Active collaborations without a contract**LightON**

Partner type: Start-up

Leader for ALMAnaCH: Djamé Seddah.

Dates: 22 Sept 2020–present

Description: LightON used to build Optical Processor Units, a specialised line of processor able to outperform GPUs on certain tasks and since the release of the Pagnol models in 2021, with the participation of Djamé Seddah, has been deploying LLMs in various industrial contexts ever since. This then informal collaboration has led to a successful BPI "communs numériques" project named Scribe where both LightOn and Almanach are co-PIs with two CIFRE PhDs starting in February 2026, both co-supervised by Djamé Seddah and Amélie Chatelain and Iacopo Poli (Lighton). They will in post-training strategies (Reinforcement learning, Reasoning models, synthetic data generation, etc.).

zaion

Participants: Chloé Clavel, Lorraine Vanel.

Partner type: Company

Leader for ALMAnaCH: Chloé Clavel.

Dates: 1 Feb 2022–1 Mar 2025

Funding received: €16,000

Description: CIFRE PhD thesis between Telecom-Paris and Zaion in order to develop conversational systems integrating socio-emotional strategies in an explicit way. The CIFRE contract being with TelecomParis, it is mentioned in this section rather than in the previous one.

10 Partnerships and cooperations**10.1 International initiatives****10.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program**

Inria Associated Team SPHERE

Participants: Justine Cassell.

Duration: 1 Jan 2025–31 Dec 2027.

PIs: Justine Cassell and Guillaume Dumas

Coordinator for ALMANACH: Justine Cassell

Partner: • MILA

Funding: €30,000

Summary: Social Physiology and Human-like Embodied Response Engineering.

10.1.2 Participation in other International Programs**Informal initiative Universal Dependencies Project**

Participants: Djamé Seddah, Benoît Sagot, Arij Riabi.

Duration: 1 Jan 2017–present.

PIs: Joakim Nivre and Christopher Manning

Coordinator for ALMANACH: Djamé Seddah

Partners: • LLF

- SEMMAGRAME
- Uppsala University
- Stanford University

Funding: €0

Summary: The Universal Dependencies project is an open community effort with over 300 contributors producing nearly 200 treebanks in over 100 languages. Universal Dependencies is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. With a release every 6 months since 2017, the UD framework is the de-facto standard for syntactic structures representations and of course the basis for most if not all supervised neural parsers.

Bilateral collaboration Interpersonality

Participants: Justine Cassell, Gabrielle Alimi, Barokshana Baskaran, Imani Stone.

Duration: 1 Jan 2024–present.

PI: Justine Cassell

Partner: • KETI (Corée)

Funding: €750

Summary: Computational models of personality have largely ignored the subtleties of expressions of personality, and have assumed that individuals (and therefore embodied conversational agents) should have one fixed personality (such as introvert or extrovert), despite the fact that all of the psychological evidence demonstrates that the expression of one's personality changes as a function of who one is speaking to, what one is speaking about, the personalities of the other individuals in a conversation, one's culture, one's age, etc. A new model is needed that takes into account the malleability of personality and therefore the need for conversational agents to also have malleable representations of personality so as to be maximally effective in working with people.

10.2 International research visitors

10.2.1 Visits of international scientists

Other international visits to the team

Marine Carpuat

Status Researcher

Institution of origin: University of Maryland

Country: United States of America

Dates: 2 Sept 2024–30 Jun 2025

Context of the visit:

Mobility programme/type of mobility: Sabbatical

Research stays abroad Hugo Scheithauer was an intern at the National Institute of Informations (NII) in Japan from 8 Apr 2024 to 30 Sept 2025.

Research stays in France Hugo Scheithauer was an intern at Naver Labs in Grenoble from 1 Sept 2024 to 28 Febr 2025.

10.3 European initiatives

10.3.1 Horizon Europe

Horizon Europe ATRIUM

Participants: Thibault Clérice, Sarah Beniere, Benjamin Kiessling, Alix Chagué.

Duration: 1 Jan 2024–31 Dec 2027.

PI: Laurent Romary

Coordinator for ALMAnaCH: Thibault Clérice

Partners:

- Institut National de Recherche en Informatique Et Automatique (Inria), France
- Archeologický ústav AV ČR, Praha v. v. i. (ARUP), Czechia
- Ludwig-Maximilians-Universität München (LMU München), Germany
- Foxcub, France
- Instytut Badań Literackich Polskiej Akademii Nauk (IBL PAN), Poland
- The University of Sheffield (USFD), United Kingdom

- Open Access in the European Area through Scholarly Communication (OPERAS), Belgium
- Stichting Radboud Universiteit, Netherlands
- Centar za digitalne humanističke nauke (Belgrade Center for Digital Humanities), Serbia
- University Of South Wales Prifysgol de Cymru (USW), United Kingdom
- Ariadne Research Infrastructure, Belgium
- Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis (Athena - Research And Innovation Center), Greece
- Idryma Technologias Kai Erevnas (Foundation For Research And Technologyhellas), Greece
- Instytut Chemii Bioorganicznej Polskiej Akademii Nauk, Poland
- Digital Research Infrastructure for the Arts and Humanities (DARIAH ERIC), France
- Laboratório Nacional De Engenharia Civil (LNEC), Portugal
- Archeologický ústav AV ČR, Brno v. v. i., Czechia
- University of York, United Kingdom
- Consiglio Nazionale delle Ricerche (CNR), Italy
- PIN Soc. Cons. A R.L. - Servizi Didattici e Scientifici per l'Università di Firenze (PIN SCRL), Italy
- Prisma Cultura S.R.L., Italy
- Clarin Eric (Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium), Netherlands
- Université de Tours, France
- Univerzita Karlova (CU), Czechia
- Athens University of Economics And Business - Research Center (AUEB-RC), Greece
- Österreichische Akademie der Wissenschaften (OeAW), Austria
- The Cyprus Institute, Cyprus
- Göteborgs universitet (UGOT), Sweden
- Net7 S.R.L., Italy

Funding: €7,000,000 total, €300,000 for ALMANACH

Summary: ATRIUM aims to empower Arts and Humanities scholars in their use of digital methods by facilitating access to a wide range of reusable workflows and interoperable, composable services offered by leading research infrastructures in the Arts and Humanities domain.

H2020 EHRI “European Holocaust Research Infrastructure”

Participants: Hugo Scheithauer, Floriane Chiffolleau, Alix Chagué, Lucas Terriel, Sarah Bénière.

Duration: 1 May 2015–28 Feb 2025.

PI: Conny Kristel (NIOD-KNAW, NL)

Coordinator for ALMANACH: Laurent Romary

Partners:

- Archives Générales du Royaume et Archives de l'État dans les provinces (Belgium)
- Aristotelio Panepistimio Thessalonikis (Greece)
- Dokumentačné Stredisko Holokaustu Občianske Združenie (Slovakia)
- Fondazione Centro Di Documentazione Ebraica Contemporanea -CDEC - ONLUS (Italy)

- International Tracing Service (Germany)
- Kazerne Dossin Memoriaal, Museum Endocumentatiecentrum Over Holocausten Mensenrechten (Belgium)
- Koninklijke Nederlandse Akademie Van Wetenschappen - KNAW (Netherlands)
- Magyarországi Zsidó Hitközsegek Szövetsége Társadalmi Szervezet (Hungary)
- Masarykův ústav a Archiv AV ČR, v. v. i. (Czech Republic)
- Memorial de La Shoah (France)
- Stiftung Zur Wissenschaftlichen Erforschung Der Zeitgeschichte - Institut Fur Zeitgeschichte IFZ (Germany)
- Stowarzyszenie Centrum Badan Nad Zaglada Zydow (Poland)
- The United States Holocaust Memorial Museum (United States)
- The Wiener Holocaust Library (UK)
- Vilniaus Gaono žydų istorijos muziejus (Lithuania)
- Wiener Wiesenthal Institut Fur Holocaust-Studien - VWI (Austria)
- Yad Vashem The Holocaust Martyrs And Heroes Remembrance Authority (Israel)
- Židovské muzeum v Praze (Czech Republic)
- Żydowski Instytut Historyczny im. Emanuela Ringelbluma (Poland)

Summary: Transforming archival research on the Holocaust, by providing methods and tools to integrate and provide access to a wide variety of archival content.

10.4 National initiatives

10.4.1 ANR

ANR MaTOS

Participants: Rachel Bawden, Éric de La Clergerie, Nicolas Dahan, Ziqian Peng, Panagiotis Tsolakis.

Duration: 1 Jan 2023–31 Dec 2026.

PI: François Yvon

Coordinator for ALMAnaCH: Rachel Bawden

Partners: • Sorbonne-Université

- Université de Paris
- CNRS

Funding: €782529 total, €280520 for ALMAnaCH

Summary: The MaTOS (Machine Translation for Open Science) project aims to develop new methods for the machine translation (MT) of complete scientific documents, as well as automatic metrics to evaluate the quality of these translations. Our main application target is the translation of scientific articles between French and English, where linguistic resources can be exploited to obtain more reliable translations, both for publication purposes and for gisting and text mining. However, efforts to improve MT of complete documents are hampered by the inability of existing automatic metrics to detect weaknesses in the systems and to identify the best ways to remedy them. The MaTOS project aims to address both of these issues.

PRME SINNet

Participants: Chloé Clavel.

Duration: 1 Mar 2024–1 Oct 2027.

PI: Chloé Clavel

Coordinator for ALMANaCH: Chloé Clavel

Funding: €474,255

Summary: SINNet proposes a paradigm shift for rendering conversational systems and social robotics a more acceptable and trustworthy technology even when using deep learning approaches. It will focus on the verbal component of the interaction, will target the agent-user social relationship, and model the behaviors indexing the state of the social relationship between agent and user, thus going beyond the analysis of the user's positive and negative sentiments. It implies developing easy-to-adapt and easy-to-explain neural models able to analyse the user's behavior contributing to user-agent co-construction processes such as the ones characterising the rapport with the agent, or the trust and affiliation in the agent, as well as to generate the agent's answer fostering the user-agent social relationship. This SINNet project will establish interdisciplinarity as a core challenge by providing a shared formalism between complex (e.g., psychological or socio-linguistic) theories of social interactions and the underlying formalism in deep learning and language models.

ANR TraLaLaM

Participants: Rachel Bawden, Benoît Sagot, Malik Marmonier.

Duration: 1 Oct 2023–30 Sept 2026.

PI: Josep Crego (Systran by ChapsVision)

Coordinator for ALMANaCH: Rachel Bawden

Partners:

- Systran by ChapsVision
- CNRS

Funding: €595,348.77 total, €169,566.82 for ALMANaCH

Summary: The aim of TraLaLaM is to explore the use of large language models (LLMs) for machine translation, by asking two main questions: (i) in what scenarios can contextual information be effectively used via prompting? and (ii) for low-resource scenarios (with a focus on dialects and regional languages), can LLMs be effectively trained without any parallel data?

ANR PRCE REVITALISE

Participants: Chloé Clavel.

Duration: 15 Feb 2022–15 Nov 2025.

PI: Magalie Ochs (LIS)

Coordinator for ALMANaCH: Chloé Clavel

Partners: • LIS

- Umanis
- ISM
- IMT Atlantique, Telecom-Paris

Funding: €580,124 total, €123,800 for ALMAnaCH

Summary: More than ever, with the increasing use of online video-conferencing solutions in daily professional interactions, public speaking skills are becoming crucial. The aim of this project is to obtain better insights into the best approaches allowing the practice of public speaking skills with technologically mediated tools. To this end, we will investigate different training environments (e.g. w/o a virtual/real audience) and different training approaches (e.g., modeling-based, feedback-based, simulation-based) to help users acquire, improve, and practice public speaking skills in full autonomy. For this purpose, different research challenges will be tackled to 1/ automatically learn, from different corpora, the multimodal cues correlated to the quality of public speaking; 2/ provide pedagogical activities rooted in coaching practice, taking a user-centered approach and 3/ provide a global evaluation of the training session as well as the specific behavioral characteristics to improve.

10.4.2 Competitiveness Clusters and Thematic Institutes

3IA PRAIRIE

Participants: Benoît Sagot, Rachel Bawden, Nathan Godey, Lydia Nishimwe, Matthieu Futral-Peter, Arij Riabi, Wissam Antoun.

Duration: 1 Oct 2019–30 Sept 2025.

PI: Isabelle Ryl

Coordinators for ALMAnaCH: Benoît Sagot, Rachel Bawden and Justine Cassell

Partners:

- Inria
- CNRS
- Institut Pasteur
- PSL
- Université de Paris
- Amazon
- Google DeepMind
- Facebook
- faurecia
- GE Healthcare
- Google
- Idemia
- Janssen
- Naver Labs
- Nokia
- Pfizer
- Stellantis

- Valeo
- Vertex

Funding: €20,000,000 total, €592,000 for ALMAnaCH

Summary: The PRAIRIE Institute (PaRis AI Research InstitutE) is one of the four French Institutes of Artificial Intelligence, which were created as part of the national French initiative on AI announced by President Emmanuel Macron on May 29, 2018. PRAIRIE’s objective is to become within five years a world leader in AI research and higher education, with an undeniable impact on economy and technology at the French, European and global levels. It brings together academic members (“PRAIRIE chairs”) who excel at research and education in both the core methodological areas and the interdisciplinary aspects of AI, and industrial members that are major actors in AI at the global level and a very strong group of international partners. Benoît Sagot and Justine Cassell hold PRAIRIE chairs. Rachel Bawden holds a junior (*tremplin*) PRAIRIE chair.

AI Cluster PRAIRIE-PSAI

Participants: Benoît Sagot, Djamé Seddah, Justine Cassell, Rachel Bawden, Chloé Clavel, Kshitij Ambilduke.

Duration: 1 Jan 2025–31 Dec 2028.

PI: Isabelle Ryl

Coordinators for ALMAnaCH: Benoît Sagot, Djamé Seddah, Justine Cassell, Rachel Bawden and Chloé Clavel

Partners:

- Inria
- CNRS
- Institut Pasteur
- PSL
- Université Paris Cité
- Google DeepMind
- META
- faurecia
- Google
- Idemia
- Janssen
- Naver Labs
- Nokia
- Pfizer
- Stellantis
- Valeo
- Vertex

Funding: €75,000,000 total

Summary: PR[AI]RIE-PSAI (Paris School of AI), a follow-up to the PRAIRIE institute, is the largest of the AI Clusters established as part of the France 2030 national strategy. This ambitious project, led by PSL University, aims to create an internationally renowned school specialising in artificial intelligence. Its goal is to advance knowledge in AI, provide world-class higher education, and produce groundbreaking innovations in this field. Benoît Sagot, Djamé Seddah and Justine Cassell hold PRAIRIE-PSAI chairs. Rachel Bawden and Chloé Clavel are PRAIRIE-PSAI fellows.

LabEx EFL

Participants: Benoît Sagot, Djamé Seddah, Éric Villemonte de La Clergerie, Virginie Mouilleron.

Duration: 1 Oct 2010–30 Sept 2024.

PI: Barbara Hemforth (LLF)

Coordinators for ALMAnaCH: Benoît Sagot, Djamé Seddah and Éric de La Clergerie

Summary: Empirical foundations of linguistics, including computational linguistics and natural language processing. ALMAnaCH’s predecessor team ALPAGE was one of the partner teams of this LabEx, which gathers a dozen of teams within and around Paris whose research interests include one aspects of linguistics or more. Several ALMAnaCH members are now “individual members” of the LabEx EFL. Benoît Sagot serves as deputy head (and former head) of one of the scientific strands of the LabEx, namely strand 6 dedicated to language resources. Benoît Sagot and Djamé Seddah are (co-)heads of a number of scientific “operations” within strands 6, 5 (“computational semantic analysis”) and 2 (“experimental grammar”). Main collaborations are related to language resource development (strands 5 and 6), syntactic and semantic parsing (strand 5, especially with LIPN [CNRS and U. Paris 13]) and computational morphology (strands 2 and 6, especially with CRLAO [CNRS and Inalco] and LLF [CNRS and Paris-Diderot]).

GDR LiLT

Participants: Benoît Sagot, Djamé Seddah, Éric de La Clergerie.

Duration: 1 Jan 2019–present.

Summary: Linguistic issues in language technology.

Duration: 1 Jan 2019–present.

Summary: Linguistic issues in language technology.

10.4.3 Other National Initiatives**Convention (MIC, Archives Nationales) NER4archives**

Participants: Cecilia Graiff.

Duration: 1 Jan 2020–27 Nov 2024.

PI: Laurent Romary

Coordinator for ALMAnaCH: Laurent Romary

Partners:

- Ministère de la culture
- Archives Nationales

Funding: €60,840

Summary: The project focuses on named entity recognition and disambiguation on data of the Archives Nationales de France (AN). The NER task is applied to the XML/EAD resources and consists in fine-tuning a spaCy based Transformer. A spaCy wrapper of the entity-fishing package is applied for entity disambiguation. Moreover, the entities are disambiguated against the Authorities made available by the AN, by leveraging RDF graph manipulation, string-matching algorithms, and an application of CrossEncoders. The idea is to merge this approach to a structure-based approach relying on GNNs, which was partially implemented.

Convention (MIC) DataCatalogue

Participants: Hugo Scheithauer, Sarah Bénière.

Duration: 12 Aug 2021–31 Oct 2024.

PI: Laurent Romary

Coordinator for ALMAnaCH: Laurent Romary

Partners:

- Ministère de la culture
- INHA
- Bibliothèque Nationale de France

Summary: The project aims at contributing to the proper transition between a basic digitalisation of cultural heritage content and the actual usage of the corresponding content within a “collection as data” perspective. To achieve this, we experiment new methods for extracting the logical structure of scanned (and OCRred) catalogues and standardise their content for publication towards curators, researchers, or wider users.

PIA project (“AMI santé numérique”) OncoLab

Participants: Éric de La Clergerie, Simon Meoni, Rian Touchent.

Duration: 1 Mar 2022–1 Mar 2026.

PI: Éric de La Clergerie

Partners:

- Arkhn
- Owkin
- Institut universitaire du cancer de Toulouse Oncopole
- Institut Curie
- Institut Bergonié
- CHU de Toulouse

Funding: €10,639,360 total, €700,720 for ALMAnaCH

Summary: The aim of the project is to make cancer data from health institutions accessible to all stakeholders involved for research and innovation purposes. The data at hand will be standardised and structured, in particular by extracting information from textual documents.

BNF Datalab PaRAMHTRS

Participants: Thibault Clérice, Rachel Bawden.

Duration: 1 Jan 2025–31 Dec 2025.

PI: Thibault Clérice

Coordinator for ALMAnaCH: Thibault Clérice

Partners:

- David Smith (Northeastern University)
- Ariane Pinche (CIHAM, CNRS)
- Gennaro Ferrante (Federico II, Naples)

Funding: €4,260

Summary: The PaRAMHTRS project advances large-scale experiments on medieval manuscripts (7th–15th centuries) in Latin and vernacular languages, leveraging HTR technology. It focuses on creating extensive corpora for culturomic studies and training models for ancient languages, while resolving abbreviations in HTR outputs. These efforts aim to enhance manuscript research and computational philology.

ExcellencES TIERED

Participants: Djamé Seddah, Benoît Sagot.

Duration: 1 Jan 2023–31 Dec 2033.

PI: SciencesPo

Coordinator for ALMAnaCH: Benoît Sagot

Partners:

- SciencesPo
- CNRS
- Ifremer
- INED
- Inserm
- Université Paris Cité
- INALCO
- IDDRI

Funding: €16,000,000 total

Summary: The ambition of the ExcellencES TIERED project is to address the challenges of democratic systems in the face of environmental transformations and the digital transition, by producing outstanding scientific research, disseminating it within society, and training today's and tomorrow's decision-makers.

Biblissima+ Grant HTRogène

Participants: Thibault Clérice, Alix Chagué.

Duration: 1 Jan 2024–31 Dec 2025.

PIs: Thibault Clérice and Alix Chagué.

Coordinators for ALMANaCH: Thibault Clérice and Alix Chagué.

Partners:

- PSL
- Ca'Foscari
- CNRS

Funding: €20,000 total

Summary: The project focuses on the production of transcriptions for literary manuscripts and public or private archives in Romance languages from the 11th to the 16th centuries. The main goal of the project is to produce training data and transcription models that are resilient to language and hand changes. HTRogenic is therefore envisaged as a building block for the infrastructure of Biblissima+ and the medieval philology of Romance languages: the project does not focus on a particular text or a small selection of texts, but on the contrary aims to produce examples of transcription capable of to constitute a representative sample. This sampling is based on specific criteria of language, script, genre and even dating.

Justine Cassell's Choose France Chair

Participants: Justine Cassell, Clara Coridon, Sinem Demirkan, Sophie Etling, Marius Le Chapelier, Reem Al Najjar, Giovanni Duca, Yassine Machta, Marie-Salomé Nsingi Kinkela Butel.

Duration: 1 Nov 2022–31 Oct 2027.

PI: Justine Cassell

Funding: €1,000,000 total

BPI project Code Commons

Participants: Benoît Sagot, Djamé Seddah.

Duration: 1 Nov 2024–31 Oct 2026.

PI: Roberto Di Cosmo

Coordinator for ALMANaCH: Djamé Seddah

Summary: CodeCommons is a two-year project building on the foundation of Software Heritage, the world's largest public source code archive. Funded by the French government with academic partners in France and Italy, our mission is to expand and enhance the archive, consolidating critical, qualified information needed to create smaller, higher-quality datasets for the next generation of responsible AI tools. It prioritizes transparency and traceability, empowering model builders and users to respect creators' rights while fostering a more sovereign and sustainable approach to AI development, massive software analysis, and reproducibility in research.

BPI project SCRIBE

Participants: Djamé Seddah, Virginie Mouilleron, Maxence Lasbordes, Melce Hüsiünbeyi.

Duration: 1 Nov 2024–31 Oct 2026.

PI: ALLONIA

Coordinator for ALMAnaCH: Djamé Seddah

Partners: • ALLONIA

- LightOn
- IDRIS/CNRS

Funding: €9,800,000 total, €1,087,440 for ALMAnaCH

Summary: The SCRIBE project is part of a strategic initiative for French digital sovereignty, aiming to develop large-scale language models (LLMs) tailored to the country's cultural, economic, and regulatory specificities. Led by a consortium of academic and industrial partners, it seeks to establish a robust, secure, and modular AI infrastructure, facilitating the industrialization and widespread adoption of LLMs. By integrating custom benchmarks, evaluation tools, and an open-source platform, SCRIBE aims to reduce reliance on foreign models while accelerating AI innovation and competitiveness in Europe.

BPI project FG4H

Participants: Éric de La Clergerie, Rian Touchent.

Duration: 1 Sept 2024–1 Sept 2026.

PI: ARKHN

Coordinator for ALMAnaCH: Éric de La Clergerie

Partners: • ARKHN

- Inria
- CHU Reims
- CHU Toulouse
- CHU Lille
- CHU Dijon-Bourgogne
- CHU Metz-Thionville
- Hopitaux universitaires Strasbourg

Funding: €4,498,882 total, €629,953 for ALMAnaCH

Summary: FG4H is a project aimed at creating a Large Language Model (LLM) for French medical data provided by a large panel of health institutions.

Detecting Dataset Manipulation and Weaponisation of NLP Models (grant)

Participants: Djamé Seddah, Benoît Sagot, Wissam Antoun.

Duration: 1 Jan 2023–31 Dec 2026.

PIs: Djamé Seddah and Benoît Sagot.

Coordinators for ALMAnaCH: Djamé Seddah and Benoît Sagot.

Partner: • Ministry of the Interior, France

Funding: €184,000

Summary: Training Large Language Models (LLMs) has become more accessible than ever due to the increased interest in scaling these LMs to obscene scales, which have been shown to not only improve performance but to unlock new emergent capabilities. However, the high compute cost required to train LLMs is exclusive to high-budget private institutions or some countries, thus raising questions about bad actors with malicious intents. Furthermore, The Center on Terrorism, Extremism, and Counter-terrorism (CTEC) highlights the upcoming threat of industrialized terrorist and extremist propaganda using models like GPT-3. Hence, it is imperative to research methods to 1) detect and defend against threats of LM weaponization and malicious dataset tampering, 2) eliminate or mitigate the threats present in language models, and 3) improve the robustness of our OSINT and threat analysis defense systems against adversarial attacks.

DEFI Inria COLaF

Participants: Benoît Sagot, Thibault Clérice, Rachel Bawden, Juliette Janès, Rasul Dent, Oriane Nédey.

Duration: 1 Aug 2023–31 Jul 2027.

PIs: Benoît Sagot and Slim Ouni.

Coordinator for ALMAnaCH: Benoît Sagot.

Partner:

- MULTISPEECH (Inria Nancy)

Funding: €1,500,000 total, €750,000 for ALMAnaCH

Summary: The Inria DEFI COLaF (Corpus and Tools for the Languages of France) aims to strengthen the ecosystem of automatic text and speech processing for the languages and speakers of France. To do this, it aims to create open datasets and use them to develop open-source models and tools.

Inria Action Exploratoire SaLM

Participants: Chloé Clavel, Benoît Sagot.

Duration: 1 Jan 2024–31 Dec 2028.

PIs: Djamé Seddah and Jean-Philippe Cointet.

Coordinator for ALMAnaCH: Djamé Seddah.

Partner:

- Sciences Po (Medialab)

Funding: €154,600 total, €130,000 for ALMAnaCH

Summary: SaLM is an interdisciplinary project between Inria Paris and Sciences Po that aims to redefine current NLP, LLM-based, algorithms by incorporating social contexts into their development and evaluation. It emphasises the importance of understanding language as a reflection of cultural and social identities. To explore this sociological dimension in NLP, the project will include two interrelated PhD projects about hate speech detection and cultural bias detection, gathering a mixed team of sociologists and NLPers to measure the role of the social dimension and prepare sociologically aware language models.

Inria Action Exploratoire BackInTime

Participants: Thibault Clérice, Hassen Aguil, Benjamin Kiessling, Benoît Sagot, Rachel Bawden.

Duration: 1 Sept 2024–31 Dec 2028.

PI: Cécile Pierrot

Coordinator for ALMAnaCH: Thibault Clérice

Partner:

- CARAMBA

Summary: BACK IN TIME brings together the expertise of researchers in three fields - artificial intelligence, cryptography and history - to decipher encrypted historical letters, some of which have lain dormant for several centuries. Given the sheer number of pages and the variety of symbols and rules involved, our aim is to develop a software package to assist or even automate the deciphering of documents from ancient, medieval and modern history.

Programme Inria Quadrant Corpus Liberatum Linguae Graecae

Participants: Thibault Clerice, Benoît Sagot, Nicolas Angleraud, Antonia Karamegkou.

Duration: 1 Mar 2025–28 Feb 2026.

PI: Thibault Clerice

Partners:

- Persée
- UMR Hisoma

Funding: €205000.0 total, €205000.0 for ALMAnaCH

Summary: The Corpus Liberatum Linguae Graecae (CLLG) aims to advance open-access corpora for Ancient Greek and improve image-to-XML OCR methodologies. CLLG will develop a scalable and efficient digitisation pipeline designed to produce a reusable, high-quality corpus fully compatible with existing scholarly resources. The project places particular emphasis on the automatic structural annotation of Ancient Greek texts, adhering to the rigorous standards and practices of Ancient Greek philology.

10.5 Regional initiatives

Domaine de recherche et d'innovation majeurs (DIM) AI4IDF

Participants: Chloé Clavel, Khaled Benaida.

Duration: 1 Sept 2021–1 Sept 2026.

PI: Chloé Clavel

Partners:

- PRAIRIE
- DataIA
- Hi!Paris
- SCAI

Summary: AI4IDF aims to deepen knowledge in AI while keeping the human being at the center of its concerns. The Paris Region must play a major role in this future sector, thanks to its scientific excellence and its incomparable ecosystem.

Domaine de recherche et d'innovation majeurs (DIM) Patrimoines matériels - innovation, expérimentation et résilience

Participants: Alix Chagué, Thibault Clérice.

Duration: 1 Jan 2022–31 Dec 2026.

Coordinator for ALMAnaCH: Laurent Romary.

Summary: The DIM Patrimoines matériels - innovation, expérimentation et résilience (PAMIR) aims to bring out new forms of social, environmental and economic development by connecting museums, companies, the Ile-de-France ecosystem of creating and crafts, universities, and laboratories around questions of fundamental and applied research on heritage collections and issues.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

- Anh Ha Ngo: Member of the organising committee for YRRSDS 2025.
- Rachel Bawden: Member of the organising committee for the WMT general shared task.

Inria-internal events

- Cecilia Graiff and Wissam Antoun: Organisers of the ALMAnaCH reading group.
- Rachel Bawden: Organiser of the ALMAnaCH seminar series.

11.1.2 Scientific events: selection

Chair of conference program committees

- Alix Chagué: Programme chair for [Digital Humanities Summer Institute's Aligned Conference](#).

Reviewer and Member of the Conference Program Committees

- Aina Garí Soler: Area chair for CoNLL and ACL Rolling Reviews. Reviewer for SRW ACL.
- Armel Zebaze: Reviewer for EMNLP 2025 and ICLR 2026.
- Chloé Clavel: Senior area chair for NAACL 2025 (Human-centered NLP Track). Reviewer for EMNLP Industry track.
- Éric Villemonte De La Clergerie: Reviewer for LREC 2026, EIPIA 2025, COLM 2025, SyntaxFest 2025 and CORIA TALN 2025.
- Hugo Scheithauer: Reviewer for DH2026.
- Justine Cassell: Reviewer for SIGDIAL, IWSDS (International Workshop on Spoken Dialogue Systems), International Conference on Multimodal Interaction (ICMI) (Doctoral consortium) and ACL Rolling Reviews.
- Lydia Nishimwe: Reviewer for WMT 2025 and LREC 2026.
- Rachel Bawden: Reviewer for ACL SRW 2025 (Student research workshop), ACL Rolling Review (February, May), EvalLLM2025 workshop, WMT 2025 and LREC 2025.
- Benoît Sagot: Reviewer for ICML 2025, ACL Rolling Reviews (February, May, July), TALN 2025, WMT 2025 and ICLR 2026.
- Djamé Seddah: Reviewer for ACL Rolling Reviews (February, July, October), LREC 2026, COLM 2025, NLP+CSS, Second Workshop on Multimodal Semantic Representations, 6th International Workshop on Designing Meaning Representations, WinNLP, SyntaxFest 2025 and CORIA-TALN 2025. Senior area chair for EMNLP 2025 (Syntax and Morphology Track) and ACL Rolling Reviews (May).
- Simon Meoni: Programme committee member for CL4Health 2025.
- Wissam Antoun: Reviewer for ArabicNLP 2025, COLING 2025 and ACL Rolling Review 2025.

11.1.3 Journal

Member of the editorial boards

- Chloé Clavel: Member of the editorial board for *IEEE Transactions of Affective Computing*, *Transactions of the Association of Computational Linguistics* (Action Editor) and *JMUI Special Issue on Socially Interactive Agents: Reinforcing or combating discrimination?* (Guest Editor - Journal on Multimodal User Interfaces).
- Rachel Bawden: Member of the editorial board for *Northern European Journal of Language Technology* and *Revue TAL* (Secretary).

Reviewer - reviewing activities

- Éric Villemonte De La Clergerie: Reviewer for *Computational Linguistics*.
- Élodie Étienne: Reviewer for *Psychology & Marketing*.
- Djamé Seddah: Reviewer for *TACL*.
- Wissam Antoun: Reviewer for *Language Resources and Evaluation* and *Neurocomputing*.

11.1.4 Invited talks

- Alix Chagué:
 - Institute of Advanced Studies, Princeton, Etats-Unis (13 Jun 2025): “HTR-United schema for dataset descriptions”.
- Chloé Clavel:
 - Saint-Denis, France (2 Apr 2025): “Best Practices”. Symposium Européen du Numérique et de l’IA Responsables (Région Ile de France)
 - AI Grid Spring School (7 Apr 2025): “Generative AI and Social Interactions”. [Website](#)
 - TALEP team, Aix-Marseille University (25 Sept 2025): “Computational Models of Socio-emotional Interactions in the Era of LLMs - the Challenges of Transparency”.
 - “Understanding social interactions in the era of LLMs: The challenges of transparency”, LATTICE, ENS (14 Oct 2025), Keynote - European Conference on Artificial Intelligence, ECAI 2025, LUHME workshop Language Understanding in the Human-Machine Era, Bologne, Italie (26 Nov 2025), and DIC/ISC/CRIA Seminar in Cognitive Informatics, Université du Québec à Montréal (UQÀM) (27 Nov 2025).
 - Webinar Matching of PEPR Ensemble (18 Dec 2025): “NLP-Driven Models of Collaboration in Human-Human and Human-Agent Interactions”.
- Éric Villemonte De La Clergerie:
 - Webinar IAS PROMESS Staff Week (online) (18 Nov 2025): “Towards more open medical datasets ?”.
 - Sofcot - 99ème Congrès de la Société Française de Chirurgie Orthopédique & Traumatologique, Paris (12 Nov 2025): “Keynote IA : enjeux de fiabilité et de recevabilité des LLM”. Joint with Jules Descamps.
- Célia Nouri:
 - Journée d’étude de l’Arcom (13 Nov 2025): “Contextualiser les interactions en ligne : Des modèles pour détecter les interventions critiques et le langage abusif”. [Website](#)
 - Seminar of the Digital Regulation Expertise Center (PEReN), Ministère de l’Économie et des Finances, Paris (1 Jul 2025): “Detecting abusive language in online conversations with graph models”.
- Carlo Santagiustina:
 - Brussels Institute for Advanced Studies (BRIAS), Brussels, Belgium. (5 Dec 2025): “Representation, Political Discontent and e-Petitions: How economic ideology and anti-elitism drive online petitioning in the EU”. [“Resilient Democracy Online: Reimagining Civic Debate on Social Media” workshop](#)
- Élodie Étienne:
 - ISIR, campus de Jussieu, Paris 5e (15 Dec 2025): “Étude des biais de genre dans la collaboration multimodale humain-agent : entre observation et simulation”. Joint with Chloé Clavel, Magalie Ochs. Journée “Intelligence Artificielle, Interactions Socio-Affectives et Éthique”
- Justine Cassell:
 - Paris, France (7 Feb 2025): “What neuroscience can teach us about AI-child interaction”. AI impact in Sciences and Humanities - AI Action Summit’s “AI, Science and Society” conference
 - Dublin, Ireland (4 Mar 2025): “AI as learning partner”. Moving from AI Safety to AI Quality
 - “Growing up in the Digital Age” Conference (4 Mar 2025): “GenAI as learning partner not oracle”.

- Kyoto, Japan (1 Oct 2025): “IA and education”. 22nd Annual Meeting of Science Technology in Society forum (STS Forum)
 - Montreal, Canada (28 Nov 2025): “Overview of PRAIRIE”. French AI Clusters
 - Montreal, Canada (28 Nov 2025): “AI and Education”. Education and AI
 - University of Oxford, London (1 Jan 2025): “Gesture Understanding and Gesture Generation with new AI models”.
- Lucence Ing:
 - CTHS, Orléans (17 Apr 2025): “Le traitement du lexique dans les premiers imprimés, entre conservation et renouvellement : étude comparative des éditions princeps du Lancelot et du Tristan en prose”. Session “Transposer, actualiser et/ou refonder : textes et autorités à l’épreuve de l’imprimé” au 149e congrès du CTHS
 - Université de Padoue, Italie (22 May 2025): “Élargir la portée d’un modèle d’annotation linguistique : problèmes et perspectives pour le français médiéval”. Journée “Testi, varietà linguistica e trattamento automatico del ’francese d’Italia”
 - ENC, PSL, Paris (11 Jun 2025): “A Study of Lexical Evolution in Medieval French: Digital Approaches to Arthurian Romances”. Journée “Digital Approaches to Pre-Modern Texts and Manuscripts”
 - Lydia Nishimwe:
 - Centrale Nantes Alumni, Paris (25 May 2025): “Table-ronde : Les Centraliennes dans les métiers de la recherche”. One of 5 panelists.
 - Rian Touchent:
 - Inist-CNRS (16 Jun 2025): “L’utilisation de corpus textuels pour l’entraînement des modèles de langage”. Journées Istex 2025
 - Benoît Sagot:
 - Awayday (“journée au vert”) of the Inria Lyon research center in Saint-Étienne (7 Jul 2025): “Apprendre les langues aux machines ?”.
 - Djamé Seddah:
 - “Preventing Language Models Weaponization: What if their Training Data were to be Compromised?”. Social Science and Generative AI: Inquiries, Instruments, Consequences (Conférence), Science Po (5 Jun 2025) and Journées INESIA, Inria, Paris (8 Aug 2025).
 - CNIT, Paris (9 Dec 2025): “Developing LLMs on academic budget, are we building the next Maginot Line? a Perspective from the GAPeron Experience”. OpenLLM Builder Days@FOST
 - Thibault Clérice:
 - King’s College of London, London, United Kingdom (20 Jan 2025): “Building a cross-lingual dataset from medieval manuscript text recognition, challenges and outcomes of CATMuS”.
 - MSHS. Lille (13 May 2025): “Collaboration et valorisation des codes et logiciels : défis, retours d’expérience et perspectives”.
 - MSHS. Lille (14 Nov 2025): “Des données de projet aux données de projets: exemples de réutilisations de données”.
 - Université de Genève, Geneva, Switzerland (16 Dec 2025): “CoMMA A Large-scale Corpus of Multilingual Medieval Archives”. Digital humanities seminar.
 - Wissam Antoun:
 - American University of Beirut, Lebanon (25 Oct 2025): “AraBERT: An enabler for Arabic NLP”.

11.1.5 Scientific expertise

- Chloé Clavel:
 - HCERES Reviewer for GIPSA-LAB.
 - Member of ANR CE 38.
- Justine Cassell:
 - Reviewer for the ERC-SYG-SH call.
 - Reviewer for ANR AAPG.
 - Reviewer for DAAD Postdoc-NeT-AI (Reviewed project proposal).
 - Member of the Conseil National du Numérique.
- Rachel Bawden:
 - Reviewer of the thematic committee for Genci projects (Reviewing projects for the allocation of computational resources).
- Benoît Sagot:
 - Member of the scientific advisory board of ERIC CLARIN.
 - Member of the Scientific Council of ARTE France.
 - Member of IA Action Summit steering committee (Paris 2025).
 - Member of Scientific Council of CNAM school on AI and digital sciences.
 - Expert of Consultation by a committee on AI-related legal matters put in place by CSPLA, the French Higher Council for Literary and Artistic Property (17/11/2025).

11.1.6 Research administration

- Rachel Bawden:
 - Member of the scientific board of the Société de Linguistique de Paris (Co-administratrice).
 - Member of the scientific board of EAMT executive committee (Co-opted member).
- Benoît Sagot:
 - Member of the scientific board of Inria Paris’s Comité des Projets (Inria Paris research centre’s Bureau du Comité des Projets).
 - Member of the scientific board of the Société de Linguistique de Paris (Co-administrateur).

11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

11.2.1 Teaching

- Rachel Bawden:
 - Master’s course (M2), Master “Mathématiques, Vision Apprentissage”, ENS Paris-Saclay, France. CM: *Speech and Language Processing*, co-organised with Chloé Clavel, Benoît Sagot, and Emmanuel Dupoux. 3hrs.
- Justine Cassell:
 - Master’s course (M2), Cerveau Social, Sorbonne Université, Paris, France. CM: *Metacognition dans l’interaction virtuelle*.
- Chloé Clavel:

- Master’s course (M2), Artificial Intelligence & Advanced Visual Computing Master, Polytechnique. CM: *Speech emotion recognition, speech synthesis and conversational systems*. 6hrs.
- Master’s course (M2), Mastère spécialisé en Intelligence Artificielle, Telecom-Paris. CM: *Introduction to Natural Language Processing*. 3hrs.
- Master’s course (M2), Master “Mathématiques, Vision Apprentissage”, ENS Paris-Saclay, France. CM: *Speech and Language Processing*, co-organised with Rachel Bawden, Benoît Sagot, and Emmanuel Dupoux. 6hrs.
- Master’s course (M2), X Datascience, Telecom-Paris. CM: *Sentiment analysis and conversational systems*. 6hrs.
- Thibault Clérice:
 - Bachelor’s Master’s and Doctorate level, TranscriboQuest Summer School, ENS Lyon. CM: *HTR for Historical Documents*, co-organised with Ariane Pinche. 18hrs.
- Aina Garí Soler:
 - Master’s course (M2), Science & Technology for Health (ST4Health), ESPCI, PSL. CM: *Artificial Intelligence Applied to Mental Health*. 21hrs.
- Cecilia Graiff:
 - Master’s course (M2), Introduction to data analysis with Python, Sciences Po Paris. CM: *Introduction to data analysis with Python*. 12hrs.
 - Master’s course (M1), Projet de statistique et science des données appliquées, ENSAE. Project supervision: *Sociodemographic debiasing influence on fake news detection*, co-organised with Gabrielle Le Bellier. 10hrs.
- Lucence Ing:
 - Master’s course (M1), Master Humanités numériques, ENC, PSL. TD/TP: *Structuration XML et XML/TEI*. 10hrs.
 - Master’s course (M1), Master Humanités numériques, Université de Rouen Normandie. TD/TP: *Encodage des données : introduction*. 15hrs.
 - Master’s course (M1), Master Humanités numériques, Université de Rouen Normandie. CM: *Introduction aux humanités numériques*. 14hrs.
 - CRBC/MSHB, Brest. Workshop on eScriptorium for HTR projects: *Initiation à eScriptorium*. 6hrs.
 - ENC, PSL, Paris. Part of the workshop “Digital Approaches to Pre-Modern Texts and Manuscripts”: *A Multilingual Medieval Aligner: from Raw Data to Aligned Witnesses*. 3hrs.
 - Biblissima, Sète. Journées des clusters 5b et 7 Biblissima+: *Aquilign, un outil d’alignement multilingue pour corpus multilingues*, coorganised with Matthias Gille Levenson, Carolina Macedo. 3hrs.
- Gabrielle Le Bellier:
 - Master’s course (M1), Projet de statistique et science des données appliquées, ENSAE. Project supervision: *Sociodemographic debiasing influence on fake news detection*, coorganised with Cecilia Graiff. 10hrs.
- Célia Nouri:
 - Master’s course (M2), Master SCIA, NLP track, EPITA, Paris. CM: *Advanced Natural Language Processing*, co-organised with Francis Kulumba, Rian Touchent. 46hrs.

- Master’s course (M2), Master Data for Business, Kedge Bordeaux. CM: *The Social Impacts of Artificial Intelligence*. 15hrs.
 - Bachelor’s course (2nd year bachelor), Collège universitaire, Sciences Po Paris. TD/TP: *Sciences et sociétés*. 15hrs.
 - Master’s course (M2), Master Public Affairs, Track Digital and Technology, Sciences Po Paris. TD/TP: *Decoding Biases in Artificial Intelligence*, co-organised with Jean-Philippe Cointet. 15hrs.
 - Master’s course (M2), Master “Mathématiques, Vision Apprentissage”, ENS Paris-Saclay, France. Project supervision for the *Speech and Language Processing* course. 4hrs.
- Benoît Sagot:
 - we are_, Paris. IA training by We Are_ school for the Casino Group Executive Board and for the Board of Directors: *Introduction au TAL et à l’IA générative*. 5hrs + 5hrs.
 - Servier headquarters, Suresnes, France. Invited talk at the AI Day of the Institut Servier de Médecine Translationnelle: *Teaching languages to machines*. 1hr.
 - Invited talk as part of the “Cycle supérieur du numérique” from IGPDE: *Apprendre les langues aux machines — 2 ans après*. 1hr.
 - Master’s course (M2), Master “Mathématiques, Vision Apprentissage”, ENS Paris-Saclay, France. CM: *Speech and Language Processing*, coorganised with Benoît Sagot, Chloé Clavel and Emmanuel Dupoux. 6hrs.
 - Carlo Santagiustina:
 - Moulin d’Andé, 65 Rue du Moulin, Andé, France. Invited to participate and contribute at the “Université d’hiver du projet RésIn”: *L’intelligence artificielle et les sciences sociales Questions, méthodes et pratiques*, co-organised with SciencesPo médialab & Université Paris Cité. 27hrs.
 - Djamé Seddah:
 - Master’s course (Executive education), Certificat Chef de projet IA, PSL-Dauphine. CM: *Natural Language Processing*. 14hrs.
 - Master’s course (M2), Master Linguistics, track Computational Linguistics, Université Paris Cité. CM: *Broadening NLP: Pretraining, instruct, alignment*. 16hrs.
 - Yi Yu:
 - Master’s course (M1), Machine learning for Text Mining, Télécom-Paris. TD/TP: *HMM*, coorganised with Matthieu Labeau and Maria Boritchev. 3hrs.
 - Bachelor’s course, Cycle Pluridisciplinaire d’Enseignement Supérieur 24/25, CPES. TD/TP: *Cycle Pluridisciplinaire d’Enseignement Supérieur 24/25*, coorganised with Maria BORITCHEV. 7hrs.
 - Master’s course (Mastères Spécialisés), IA717 Natural Language Processing 2025/2026, Télécom-Paris. Project supervision: *IA717 Natural Language Processing 2025/2026*, coorganised with Matthieu Labeau and Maria Boritchev. 18hrs.
 - Master’s course (Mastères Spécialisés), CSC_5DS25_TP Natural Language Processing and Sentiment Analysis 2025/2026, Institut Polytechnique de Paris. TD/TP: *Natural Language Processing*, coorganised with Matthieu Labeau. 7hrs.
 - You Zuo:
 - Master’s course (M1), Master pluriTAL, INALCO, France. CM: *NLP in English*. 52hrs.

11.2.2 Supervision

PhD

- Arij Riabi: “NLP for low-resource, non-standardised language varieties, especially North-African dialectal Arabic written in Latin script” (1 Oct 2021–31 Mar 2025). Supervised by Laurent Romary and Djamé Seddah. PhD defended on 18 Mar 2025.
- Lydia Nishimwe: “Robust Neural Machine Translation” (1 Oct 2021–30 Jun 2025). Supervised by Benoît Sagot and Rachel Bawden. PhD defended on 18 Jun 2025.
- Matthieu Futral-Peter: “Text-image multimodal models” (1 Nov 2021–30 Jul 2025). Inria. Supervised by Cordelia Schmid, Benoît Sagot and Rachel Bawden. PhD defended on 9 Dec 2025.
- Alix Chagué: “Methodology for the creation of training data and the application of handwritten text recognition to the Humanities.” (1 Nov 2021–present). Secondary affiliation: Université de Montréal and CRIHN. Supervised by Laurent Romary, Emmanuel Château-Dutier and Michael Sinatra.
- Alisa Barkar: “Interpretable textual features, public speeches, multimodal systems” (1 Nov 2022–1 Nov 2025). Primary affiliation: Télécom Paris. Supervised by Chloé Clavel, Beatrice Biancardi and Mathieu Chollet. PhD defended on 10 Dec 2025.
- Francis Kulumba: “Authorship attribution and verification through learned representations of writing style.” (1 Nov 2022–present). Supervised by Laurent Romary and Guillaume Vimont.
- Rian Touchent: “Information Extraction on French Electronic Health Records” (1 Dec 2022–present). Supervised by Laurent Romary and Éric Villemonte De La Clergerie.
- Simon Meoni: “Exploration of adaptation methods for neural models in the French clinical domain” (1 Dec 2022–30 Nov 2025). CIFRE PhD with Arkhn. Supervised by Laurent Romary and Éric Villemonte De La Clergerie.
- Wissam Antoun: “Detecting Dataset Manipulation and Weaponisation of NLP Models” (1 Mar 2023–present). Supervised by Benoît Sagot and Djamé Seddah.
- You Zuo: “Patent representation learning for innovation generation and technical trend analysis” (1 Mar 2023–present). CIFRE PhD with qatent. Supervised by Benoît Sagot, Éric Villemonte De La Clergerie and Kim Gerdes (CIFRE advisor).
- Yanzhu Guo: “Language model evaluation, argument mining, computational social science” (1 Oct 2023–31 Mar 2025). Primary affiliation: Ecole Polytechnique. Supervised by Michalis Vazirgiannis and Chloé Clavel. PhD defended on 20 Jun 2025.
- Lorraine Vanel: “Conversational AI, Social/emotional Dialogue Generation” (1 Oct 2023–1 Feb 2025). Primary affiliation: Télécom Paris. CIFRE PhD with Zaion. Supervised by Chloé Clavel and Alya Yacoubi (CIFRE advisor). PhD defended on 8 Jul 2025.
- Nicolas Dahan: “Evaluation of the machine translation of scientific documents” (1 Oct 2023–present). Secondary affiliation: CNRS/ISIR. Supervised by François Yvon and Rachel Bawden.
- Ziqian Peng: “Machine translation of scientific documents” (1 Oct 2023–present). Primary affiliation: CNRS/ISIR. Supervised by François Yvon and Rachel Bawden.
- Lucie Chenain: “Speech Emotion Recognition for Huntington’s Disease risky behaviour” (1 Oct 2023–present). Primary affiliation: Université Paris Cité. Supervised by Anne-Catherine Bachoud Levi and Chloé Clavel.
- Biswesh Mohapatra: “Improving chatbot dialogue systems through collaborative grounding” (1 Oct 2023–present). Supervised by Justine Cassell and Laurent Romary.

- Hugo Scheithauer: “Acquisition, integration and redistribution of structured data in GLAMs: harmonising practices” (1 Nov 2023–present). Supervised by Laurent Romary, Thibault Clérice and Gonzalo Romero-García.
- Armel Zebaze: “Analogy for Multilingual Natural Language Processing” (1 Nov 2023–present). Supervised by Benoît Sagot and Rachel Bawden.
- Rasul Dent: “Large-scale language identification (numerous languages, massive data, distinction between closely related varieties) with a focus on the languages of France and French-based creoles.” (1 Nov 2023–present). Supervised by Benoît Sagot, Thibault Clérice and Pedro Ortiz.
- Anh Ha Ngo: “Multimodal models, conversation repair and human-agent interaction” (1 Jan 2024–present). Secondary affiliation: Sorbonne Université. Supervised by Chloé Clavel and Catherine Pelachaud, Nicolas Rollet.
- Pierre Chambon: “Code generation with language models” (26 Feb 2024–present). Supervised by Benoît Sagot and Gabriel Synnaeve (CIFRE advisor).
- Oriane Nédey: “Machine Translation for low-resource dialectal variants” (1 Oct 2024–present). Supervised by Benoît Sagot, Rachel Bawden and Thibault Clérice.
- Reem Al Najjar: “Investigating Neural Mechanisms of Collaboration Among Peers” (1 Nov 2024–present). Supervised by Justine Cassell.
- Gabrielle Le Bellier: “Controlled generation for bias mitigation and cultural awareness in conversational language models” (1 Nov 2024–present). Supervised by Benoît Sagot and Chloé Clavel.
- Célia Nouri: “Toxicity and Opinions Detection and Analysis on Social Media Conversations” (1 Nov 2024–present). Secondary affiliation: Sciences Po médialab. Supervised by Chloé Clavel and Jean-Philippe Cointet (médialab).
- Cecilia Graiff: “Multilingual and cross-cultural automatic analysis of argumentation structures in political debates” (1 Dec 2024–present). Supervised by Benoît Sagot and Chloé Clavel.
- Yi Yu: “Automatic analysis of the human ability to collaborate in dyadic and group conversations, with a view to educational applications.” (1 Dec 2024–present). Secondary affiliation: Telecom-Paris. Supervised by Chloé Clavel and Maria Boritchev (Telecom-Paris).
- Xiangyu An: “Evaluation criteria and optimization strategies for socio-conversational systems” (2 Jun 2025–present). CIFRE PhD with Afnor. Supervised by Chloé Clavel and Abdallah Essa (Afnor) (CIFRE advisor).
- Romain Froger: “Planning and Reasoning in Dynamic Environments: Towards Real-Time, Interactive, Intelligent, and Efficient Systems with Large Language Model Agents” (16 Jun 2025–present). Primary affiliation: META AI Paris. CIFRE PhD with META. Supervised by Djamé Seddah and Thomas Scialom (Meta) (CIFRE advisor).
- Oussama Silem: “Rapport-Building in LLM-Powered Conversational Agents” (1 Sept 2025–present). Supervised by Justine Cassell.
- Sinem Demirkan: “Neural Representations of Multilingual Language Processing” (1 Sept 2025–present). Supervised by Benoît Sagot.
- Sofia De Tremiolles: “Socio-cultural biases” (1 Nov 2025–present). Secondary affiliation: medialab. Supervised by Djamé Seddah and Jean-Phillipe Cointet.

Interns

- Imani Stone: “Hyperscanning of pairs of children” (1 Jan 2025–30 Jun 2025). Supervised by Justine Cassell.
- Zofia Milczarek: “Prompting and fine-tuning LLMs for natural speech-like (text) dialogue generation” (20 Jan 2025–20 Jul 2025). Supervised by Justine Cassell and Marius Le Chapelier.
- Gabrielle Alimi: “Hyperscanning of pairs of children” (27 Jan 2025–27 Jun 2025). Supervised by Justine Cassell.
- Barokshana Baskaran: “Hyperscanning of pairs of children” (27 Jan 2025–27 Jun 2025). Supervised by Justine Cassell.
- Clara Coridon: “How rapport is built in French conversations through verbal and nonverbal behaviors” (3 Mar 2025–3 Sept 2025). Supervised by Justine Cassell.
- Giovanni Duca: “Reinforcement learning techniques to improve conversational grounding in dialogue systems” (3 Mar 2025–31 Jul 2025). Supervised by Justine Cassell and Biswesh Mohapatra.
- Théo Lasnier: “Code generation evaluation” (1 Apr 2025–30 Sept 2025). Supervised by Djamé Seddah.
- Adriano Rivierez: (1 May 2025–29 Aug 2025). Supervised by Chloé Clavel and Anh Ha Ngo.
- Kshitij Ambilduke: “Exploring multimodal interactions for video game corpus” (5 May 2025–29 Aug 2025). Supervised by Djamé Seddah.
- Sofia Imbert De Tremiolles: “Socio Cultural biases” (5 May 2025–29 Aug 2025). Primary affiliation: Medialab Science Po. Supervised by Djamé Seddah and Jean-Phillipe Cointet.
- Yassine Machta: “Non-verbal behavior and gesture generation for enhancing the believability and expressiveness of virtual agents.” (15 May 2025–21 Nov 2025). Supervised by Justine Cassell.
- Mayank Palan: “Reinforcement learning techniques to improve conversational grounding in dialogue systems” (19 May 2025–14 Aug 2025). Supervised by Justine Cassell.

Engineers

- Virginie Moulleron: “Correction and annotation of the Alien vs Predator dataset, Prompt Tuning and Data extraction from LLMs.” (1 Dec 2022–present). Supervised by Djamé Seddah.
- Juliette Janès: “Recovery, encoding, maintenance, and publication of textual data on French and other languages of France produced within the framework of the DEFI COLaF” (1 Oct 2023–present). Supervised by Benoît Sagot and Thibault Clérice.
- Sarah Bénière: “Automatic analysis of digitized sales catalogs” (1 Oct 2023–31 Jan 2025). Supervised by Laurent Romary.
- Samuel Scalbert: “Detection of software in HAL articles using GROBID and Softcite in the context of the GrapOS project.” (1 Oct 2023–31 Mar 2025). Supervised by Laurent Romary.
- Marius Le Chapelier: “Developing the SARA (Socially Aware Robot Assistant) dialogue system to be able to build social bonds (rapport) with users in order to improve performance.” (1 Nov 2023–present). Supervised by Justine Cassell.
- Sinem Demirkan: “Hyperscanning of pairs of children in order to better understand the neural correlates of rapport” (1 Jan 2024–2 Jan 2025). Supervised by Justine Cassell.
- Malik Marmonier: “Machine Translation with large language models in low-resource scenarios and for unseen languages” (1 May 2024–present). Supervised by Rachel Bawden and Benoît Sagot.

- Cindy Evellyn de Araujo Silva: “Hyperscanning of pairs of children in order to better understand the neural correlates of rapport, and the impact of rapport on learning” (1 Jun 2024–30 Nov 2025). Supervised by Justine Cassell.
- Hasan Onur Keles: “Hyperscanning of pairs of children in order to better understand the neural correlates of rapport, and the impact of rapport on learning” (1 Jul 2024–30 Jun 2025). Supervised by Justine Cassell.
- Hassen Aguil: “Interface and back-end for automatic recognition of standard and non-standard handwriting” (1 Sept 2024–present). Supervised by Thibault Clérice.
- Sophie Etling: “Hyperscanning of pairs of children in order to better understand the neural correlates of rapport, and the impact of rapport on learning” (1 Sept 2024–present). Supervised by Justine Cassell.
- Panagiotis Tsolakis: “Scientific article management infrastructure for translation” (1 Oct 2024–present). Supervised by Rachel Bawden and Laurent Romary.
- Sinem Demirkan: “Hyperscanning of pairs of children in order to better understand the neural correlates of rapport” (1 Jan 2025–31 Aug 2025). Supervised by Justine Cassell.
- Oussama Silem: “To develop the ability of LLMs to manage both task-related interactions and social interactions that improve task performance. ” (1 Jan 2025–31 Aug 2025). Supervised by Justine Cassell.
- Lucence Ing: “Coordination and development of multilingual and multi-variety text corpora for French and the languages of France for digital humanities, computational linguistics, and NLP research.” (1 Feb 2025–present). Supervised by Benoît Sagot.
- Sarah Bénière: “Layout Analysis at scale to produce structured documents.” (1 Feb 2025–31 Dec 2025). Supervised by Thibault Clérice.
- Nathan Godey: “Continuation of work on the training and evaluation of Gaperon language models.” (1 Feb 2025–30 Apr 2025). Supervised by Benoît Sagot.
- Nicolas Angleraud: “End-to-end layout-aware and OCR-based transformation of Ancient Greek printed editions into structured TEI-XML corpora for open digital humanities research.” (1 Mar 2025–present). Supervised by Thibault Clérice and Benoît Sagot.
- Melce Hüsiünbeyi: “Development of multilingual and multimodal fact checking benchmark” (1 May 2025–31 Oct 2025). Supervised by Djamé Seddah.
- Marie-Salomé Nsingi Kinkela Butel: “Turn taking prediction in conversational agent ” (1 Jun 2025–present). Supervised by Justine Cassell.
- Benjamin Kiessling: “Ancient Cryptography Text Recognition; Layout analysis of printed document.” (1 Jul 2025–present). Supervised by Thibault Clérice.
- Khaled Benaida: (1 Oct 2025–present). Supervised by Chloé Clavel.
- Barokshana Baskaran: “Hyperscanning of pairs of children in order to better understand the neural correlates of rapport” (1 Oct 2025–present). Supervised by Justine Cassell.
- Théo Lasnier: “Code generation evaluation ; interpretability ; llm safety” (1 Oct 2025–present). Supervised by Djamé Seddah.
- Antonia Karamolegkou: “Creating annotated datasets and vision-language architectures for OCR of ancient and medieval Greek sources” (1 Nov 2025–present). Supervised by Thibault Clérice.
- Maxence Lasbordes: “Post training strategies and alignment” (1 Nov 2025–present). Supervised by Djamé Seddah.

Postdocs

- Aina Garí Soler: “Automatic analysis of alignment between speakers in conversations” (1 Oct 2024–30 Sept 2025). Secondary affiliation: Telecom-Paris. Supervised by Chloé Clavel.
- Rémy Ben Messaoud: “Hyperscanning of pairs of children in order to better understand the neural correlates of rapport” (1 Sept 2025–present). Supervised by Justine Cassell.
- Élodie Étienne: “Mitigating bias and enhancing explainability during collaboration between humans and socially interactive agents groups” (1 Oct 2025–present). Supervised by Chloé Clavel.
- Yannis Karmim: “Reducing social and cultural biases in language models using Graph structured data” (1 Nov 2025–present). Secondary affiliation: Inria Chile. Supervised by Djamé Seddah.
- Erinda Morina: “Hyperscanning of pairs of children in order to better understand the neural correlates of rapport” (1 Nov 2025–present). Supervised by Justine Cassell.

11.2.3 Juries

PhD

- Éric Villemonte De La Clergerie
 - Examiner
 - * Aman Sinha (Université de Lorraine). Title: *Evaluation of Medical Language Models*.
 - * Kun Zhang (Institut Polytechnique de Paris). Title: *Contributions to Evaluating and Improving the Faithfulness for Text Generation*.
- Rachel Bawden
 - Examiner
 - * Léane Jourdan (Nantes Université). Title: *Automatic Text Revision of Scientific Writing Assistance*.
 - PhD co-supervisor
 - * Lydia Nishimwe (Inria). Title: *Robust Neural Machine Translation of User-Generated Content*.
 - * Matthieu Futral-Peter at Inria on 9 Dec 2025. Title: *Multilingual and Multimodal Language Modelling*.
- Benoît Sagot
 - Reviewer
 - * Eliot Maës (Aix Marseille Université). Title: *Multimodal Learning and modelling information exchanges in natural interactions*.
 - PhD director
 - * Lydia Nishimwe (Inria). Title: *Robust Neural Machine Translation of User-Generated Content*.
 - * Matthieu Futral-Peter at Inria on 9 Dec 2025. Title: *Multilingual and Multimodal Language Modelling*.
- Djamé Seddah
 - Reviewer
 - * Julie Tytgat (Université Paris Cités). Title: *Entre similarité de surface et similarité sémantique: une application à la surveillance de marques*.
- Chloé Clavel

- PhD director
 - * Yanzhu Guo (Ecole Polytechnique). Title: *Automatic Evaluation of Human-Written and Machine-Generated Text*.
 - * Lorraine Vanel (Sorbonne Université). Title: *Planning Socio-Emotional Response Generation for Conversational Agents*.
 - * Alisa Barkar at Telecom-Paris on 10 Dec 2025. Title: *Automatically Interpreting LLM Judgments Using Linguistic Insights: Case of Public Speaking*.
- Examiner
 - * Lucie Galland (Sorbonne Université). Title: *A dialogue manager for dual-level adaptation to context and users' profile*.
 - * Alexis Plaquet at (IRIT, Toulouse). Title: *Contributions à l'entraînement du modèle neuronal de segmentation en locuteurs et son impact sur leur regroupement*.

Master

- Éric Villemonte De La Clergerie
 - Examiner for Matthieu Boyer (ENS Ulm). Title: *Formalizing a Functional Programming Language for Natural Language Semantics*.
 - Co-supervisor for Anh Thu Vu (Sorbonne Université). Title: *Automatic Information Extraction for Patient Recruitment in Oncology Clinical Trials and ICD Coding*.

HdR

- Chloé Clavel
 - Examiner for Brian Ravenet (LISN). Title: *Interactions Humain-Machine Socio-affectives pour Environnements Virtuels Collaboratifs*.

CSD

- Éric Villemonte De La Clergerie
 - Cyril Bruneau (Université Paris Nanterre). Title: *Transmettre des valeurs à l'école: développement d'un outillage informatique appliqué aux manuels scolaires d'histoire (1870-2020)*.
 - Jules Descamps (Université Paris Cité). Title: *An assessment of natural language processing models for enhancing health data collection from medical literature*.
 - Clément Dauvilliers (Sorbonne Université). Title: *Apprentissage automatique pour la prédiction d'événements météorologiques extrêmes*.
 - Victor Morand (Sorbonne Université). Title: *Adaptation de domaines et de langages dans les PLM, application à la recherche d'information conversationnelle*.
 - Pierre Epron (Université Paris Cité). Title: *Knowledge based Hallucination Detection & Mitigation in Large Language Models*.
 - Franck Signe Talla (Sorbonne Université). Title: *Développement de réseaux de neurones modulaires, avec application aux modèles de langue multilingues*.
- Rachel Bawden
 - Tom Calamai (Inria). Title: *Détection automatique d'argument fallacieux*.
 - Zineddine Tighidet (BNP Paribas, Sorbonne Université, CNRS). Title: *Etude de l'impact différentiel des choix de modélisation lors du développement d'un modèle de langage bancaire et application de mesures de compatibilité sémantique en tant que garde-fou pour la génération du langage*.

- Estelle Zheng (Université de Lorraine). Title: *Affinage des grands modèles de langage pour la planification et l'action via des APIs*.
- Bastien Michel (Inria). Title: *Optimization of Symmetric Cryptanalysis*.
- Maxime Poli (ENS). Title: *Pré-entraînement multilingue universel auto-supervisé de modèles de langage parlé*.
- Chloé Clavel
 - Samy Haffoudi (Telecom-Paris). Title: *Combining language models and knowledge bases: towards reliable translation of natural language into structured queries*.
 - Léo Labat (Sorbonne Université). Title: *What Are Multilingual LLMs' Values, if Any ?*.
- Benoît Sagot
 - Lingyun Gao (Université catholique de Louvain). Title: *C-mesure : Une modélisation plus fiable et adaptative de la lisibilité pour le français langue étrangère*.

Hiring committees

- Rachel Bawden:
 - Member of the *Commission des emplois scientifiques (CES)* hiring committee at Inria (Paris Centre). Delegations, postdocs and PhDs.

11.2.4 Educational and pedagogical outreach

- Juliette Janès gave a talk at (Speed meeting between students and scientists), “Filles et Maths - Rendez-vous des jeunes mathématiciennes et informaticiennes (RJMI)”. Inria Paris, 23 Oct 2025.
- Chloé Clavel gave a talk at Lycée Louis Le Grand (Présentation orale lycée, lors de la remise des prix Olympiades de mathématiques), “La reconnaissance automatique des émotions dans la voix”. , 25 Jun 2025.
- Élodie Étienne gave a talk at Lycée Paul Bert (Paris) (Inria "Chiche !" programme), “Les maths comme passeport pour les technologies du futur”. Lycée Paul Bert (Paris), 4 Dec 2025.
- Justine Cassell gave a talk at Ecole Jeannine Manuel (IAckathon - Edition collègue), “IAckathon - Edition collègue”. Paris, France, 19 Jun 2025.
- Djamé Seddah:
 - gave a talk at Association Ro-BOTs (Ecole Bachelet, 14 rue Alexandre Bachelet, Saint Ouen) (Intervention collégien 3e), Saint Ouen, 21 Oct 2025.
 - gave a talk at locaux de l'association RoBoTic, école Bachelet (Saint Ouen) (Témoignage métier - assoc. RoBoTIC), “Mon parcours et l'intelligence artificielle”. Ecole Bachelet (Saint Ouen), 21 Oct 2025.
- Gabrielle Le Bellier:
 - gave a talk at (Speed meeting between students and scientists), “Filles et Maths - Rendez-vous des jeunes mathématiciennes et informaticiennes (RJMI)”. Inria Paris, 23 Oct 2025.
 - gave a talk at (Inria “Chiche !" programme - Chiche dans les mur”), “Mon parcours et les biais dans les modèles de langue”. Inria Paris , 11 Apr 2025.
- Lydia Nishimwe:
 - gave five talks as part of the Inria "Chiche !" programme “La traduction automatique + Mon parcours”, Lycée Maurice Ravel, Lycées Jacques Prévert et Marguerite Yourcenar, Lycée de l'Essouriau, Lycée privé Charles de Foucauld, Lycée Molière
 - animated an event Journée Citoyenne IA (Interactive Demonstration), “L'apprenti illustrateur”. Hôtel de Ville de Paris, 1 Jan 2025-1 Feb 2025.

11.3 Popularization

11.3.1 Productions (articles, videos, podcasts, serious games, ...)

Articles with citation

- Rachel Bawden and Benoît Sagot cited in an article by Science & Vie Junior (432) (Media article), “Le langage c’est ta force”. 1 Sept 2025-1 Oct 2025.
- Chloé Clavel:
 - cited in an article by La Recherche (582) (Media article), “Les robots peuvent-ils rêver d’amour ?”. 1 Jul 2025.
 - cited in an article by Le Monde (Media article), “Les “hallucinations”, ces erreurs de l’IA qui rendent la machine trop humaine”. 19 Nov 2025.

11.3.2 Participation in Live events

Media interviews

- Célia Nouri interviewed as part of The Conversation (The Conversation), “Détection automatique et cyberviolences.” Inria Paris, 9 Sept 2025.
- Chloé Clavel:
 - interviewed as part of podcast Ex Machina, l’ère des algorithmes, “Les agents autonomes sont-ils la vraie révolution de l’IA générative ?” PSL, 14 Nov 2025.
 - interviewed as part of La Science au labo (CQFD), France Culture (Radio France), “Pourrait-on suivre l’état émotionnel de certains malades grâce à des IA en partie transparentes ?” , 4 Dec 2025.

Talks and panel sessions

- Justine Cassell:
 - gave two talks at Fondation de France, Timisoara, Romania (“La Nuit de la Philosophie”), “AI as partner and not oracle: a democracy-oriented approach to AI and education” and “supporting critical thinking about AI - and with AI - in order to strengthen a democratic society”. 13 and 14 Jun 2025.
 - gave a talk at Ecole Jeannine Manuel to high school students, “Can a virtual peer be a good learning partner?” Paris, 19 Jun 2025.
- Wissam Antoun participated in the panel session “L’intelligence artificielle méditerranéenne : enjeux d’ancrage, de pluralité et de responsabilité” at the Forum méditerranéen de l’intelligence artificielle 2025. Tunis, Tunisia, 21 Nov 2025.

12 Scientific production

12.1 Major publications

- [1] T. Clérice, S. Gabay, M. Vlachou-Efstathiou, A. Pinche and B. Sagot. *CoMMA, a Large-scale Corpus of Multilingual Medieval Archives*. 9th Dec. 2025. URL: <https://inria.hal.science/hal-05299220>.
- [2] P.-A. Duquenne, H. Schwenk and B. Sagot. ‘Modular Speech-to-Text Translation for Zero-Shot Cross-Modal Transfer’. In: *Proceedings of INTERSPEECH 2023*. INTERSPEECH 2023. Dublin, Ireland, 20th Aug. 2023. DOI: [10.21437/Interspeech.2023-2484](https://doi.org/10.21437/Interspeech.2023-2484). URL: <https://hal.science/hal-04264023>.

- [3] M. Futral, C. Schmid, I. Laptev, B. Sagot and R. Bawden. ‘Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 3rd July 2023, pp. 5394–5413. DOI: [10.18653/v1/2023.acl-long.295](https://doi.org/10.18653/v1/2023.acl-long.295). URL: <https://inria.hal.science/hal-03977982>.
- [4] N. Godey, W. Antoun, R. Touchent, R. Bawden, É. de la Clergerie, B. Sagot and D. Seddah. *Gaperon: A Peppered English-French Generative Language Model Suite*. 29th Oct. 2025. URL: <https://hal.science/hal-05410121>.
- [5] N. Godey, E. Villemonte de La Clergerie and B. Sagot. ‘Anisotropy Is Inherent to Self-Attention in Transformers’. In: *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). St Julians, Malta, 17th Mar. 2024, pp. 35–48. URL: <https://hal.science/hal-04593391>.
- [6] Y. Guo, G. Shang, M. Vazirgiannis and C. Clavel. ‘The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text’. In: *NAACL 2024 Findings - Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Mexico City, Mexico, 16th Apr. 2024. URL: <https://hal.science/hal-04593399>.
- [7] B. Haddow, R. Bawden, A. V. Miceli Barone, J. Helcl and A. Birch. ‘Survey of Low-Resource Machine Translation’. In: *Computational Linguistics* 48.3 (2022), pp. 673–732. URL: <https://inria.hal.science/hal-03479757>.
- [8] G. Jawahar, B. Sagot and D. Seddah. ‘What does BERT learn about the structure of language?’ In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, July 2019. URL: <https://hal.inria.fr/hal-02131630>.
- [9] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. Villemonte de La Clergerie, D. Seddah and B. Sagot. ‘CamemBERT: a Tasty French Language Model’. In: *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*. Seattle / Virtual, United States, July 2020. DOI: [10.18653/v1/2020.acl-main.645](https://doi.org/10.18653/v1/2020.acl-main.645). URL: <https://hal.inria.fr/hal-02889805>.
- [10] B. Muller, A. Anastasopoulos, B. Sagot and D. Seddah. ‘When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models’. In: *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Mexico City, Mexico, 6th June 2021. URL: <https://inria.hal.science/hal-03251105>.
- [11] C. Nouri, J.-P. Cointet and C. Clavel. ‘Graphically Speaking: Unmasking Abuse in Social Media with Conversation Insights’. In: *63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*. Vienna, Austria, 27th July 2025. URL: <https://sciencespo.hal.science/hal-05165879>.
- [12] A. Pinche, T. Clérice, A. Chagué, J.-B. Camps, M. Vlachou-Efstathiou, M. Gille Levenson, O. Brisville-Fertin, F. Boschetti, F. Fischer, M. Gervers, A. Boutreux, A. Manton, S. Gabay, W. Haverals, M. Kestemont, C. Vandyck and P. O’Connor. ‘CATMuS-Medieval: Consistent Approaches to Transcribing Manuscripts: A generalized set of guidelines and models for Latin scripts from Middle Ages (8th–16th century)’. In: *Digital Humanities - DH2024*. Washington DC, United States, 5th Aug. 2023. URL: <https://inria.hal.science/hal-04346939>.
- [13] A. Riabi, V. Mouilleron, M. Mahamdi, W. Antoun and D. Seddah. ‘Beyond Dataset Creation: Critical View of Annotation Variation and Bias Probing of a Dataset for Online Radical Content Detection’. In: *COLING 2025 - 31st International Conference on Computational Linguistics*. Abu Dhabi, United Arab Emirates, 19th Jan. 2025. URL: <https://hal.science/hal-04867863>.

- [14] A. R. Zebaze, B. Sagot and R. Bawden. ‘Compositional Translation: A Novel LLM-based Approach for Low-resource Machine Translation’. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 6th Mar. 2025, pp. 22328–22357. doi: [10.18653/v1/2025.findings-emnlp.1216](https://doi.org/10.18653/v1/2025.findings-emnlp.1216). URL: <https://hal.science/hal-05009363>.

12.2 Publications of the year

International journals

- [15] L. Chenain, A. Fabre, H. Titeux, G. Morgado, K. Youssef, C. Clavel and A.-C. Bachoud-Lévi. ‘Emotional speech markers of psychiatric disturbance in Huntington’s disease’. In: *Frontiers in Psychiatry* 16 (12th Aug. 2025), 1633492-1:1633492–20. doi: [10.3389/fpsy.2025.1633492](https://doi.org/10.3389/fpsy.2025.1633492). URL: <https://inserm.hal.science/inserm-05304543> (cit. on p. 39).
- [16] S. Gabay, A. Pinche, P. Nahon, A. Chagué, P. Jacsont, É. Paupe, J.-C. Rebetez, M. Humeau, C. Payot, T. Maillard, Y. Jauregui, E. Leblanc and L. Chappuis. ‘Reading Before It Can Be Read: Philological Reflections on Automatic Text Recognition for Modern French Manuscripts’. In: *Humanités numériques* 12 (24th Dec. 2025). doi: [10.4000/15ick](https://doi.org/10.4000/15ick). URL: <https://hal.science/hal-05431021> (cit. on p. 43).
- [17] Y. Guo, G. Shang and C. Clavel. ‘Benchmarking Linguistic Diversity of Large Language Models’. In: *Transactions of the Association for Computational Linguistics* 13 (13th Nov. 2025), pp. 1507–1526. doi: [10.1162/TACL.a.47](https://doi.org/10.1162/TACL.a.47). URL: <https://inria.hal.science/hal-05444994> (cit. on p. 28).
- [18] E. Kempf, A. Vu, E. Villemonte de La Clergerie and R. Flicoteaux. ‘415P How to generate open source annotated cancer clinical datasets with LLMs to support the development of smaller language models’. In: *ESMO Real World Data and Digital Oncology* 10 (Nov. 2025), p. 100611. doi: [10.1016/j.esmorw.2025.100611](https://doi.org/10.1016/j.esmorw.2025.100611). URL: <https://hal.science/hal-05444771>.
- [19] T. A. Nguyen, B. Muller, B. Yu, M. R. Costa-Jussa, M. Elbayad, S. Popuri, C. Ropers, P.-A. Duquenne, R. Algayres, R. Mavlyutov, I. Gat, M. Williamson, G. Synnaeve, J. Pino, B. Sagot and E. Dupoux. ‘SpiRit-LM: Interleaved Spoken and Written Language Model’. In: *Transactions of the Association for Computational Linguistics* 13 (7th Jan. 2025), pp. 30–52. URL: <https://inria.hal.science/hal-04449905> (cit. on p. 26).
- [20] A. Pinche and T. Clérice. ‘Reconnaissance automatique d’écriture et sources historiques: Limites et nouvelles perspectives’. In: *Studi francesi* 206 (2025), pp. 357–366. URL: <https://shs.hal.science/halshs-05300912> (cit. on p. 43).

Invited conferences

- [21] A. Chagué. ‘HTR-United schema for dataset description’. In: 2025 Workshop SCOOP - Source Codes of the Past. Princeton, NJ, United States, 12th June 2025. URL: <https://inria.hal.science/hal-05117083> (cit. on p. 43).

International peer-reviewed conferences

- [22] J. O. Alabi, I. A. Azime, M. Zhang, C. España-Bonet, R. Bawden, D. Zhu, D. I. Adelani, C. O. Odoje, I. Akinade, I. Maab, D. David, S. H. Muhammad, N. Putini, D. Ademuyiwa, A. Caines and D. Klakow. ‘AFRIDOC-MT: Document-level MT Corpus for African Languages’. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 10th Jan. 2025, pp. 27758–27794. doi: [10.18653/v1/2025.emnlp-main.1413](https://doi.org/10.18653/v1/2025.emnlp-main.1413). URL: <https://inria.hal.science/hal-04903006> (cit. on p. 29).

- [23] J. Alejandro Lopetegui Gonzalez, A. Riabi and D. Seddah. ‘Common Ground, Diverse Roots: The Difficulty of Classifying Common Examples in Spanish Varieties’. In: *VarDial 2025 - Twelfth Workshop on NLP for Similar Languages, Varieties and Dialects co-located with COLING 2025*. Abu Dhabi, United Arab Emirates, 19th Jan. 2025. URL: <https://hal.science/hal-04868010> (cit. on p. 34).
- [24] W. Antoun, B. Sagot and D. Seddah. ‘ModernBERT or DeBERTaV3? Examining Architecture and Data Influence on Transformer Encoder Models Performance’. In: *IJCNLP-AAACL 2025 - 14th International Joint Conference on Natural Language Processing (IJCNLP) and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AAACL)*. Mumbai, India, 14th Nov. 2025. URL: <https://hal.science/hal-05365631> (cit. on p. 26).
- [25] A. Barkar, M. Chollet, M. Labeau, B. Biancardi and C. Clavel. ‘Decoding Persuasiveness in Eloquence Competitions: An Investigation into the LLM’s Ability to Assess Public Speaking’. In: *ICAART 2025 - 17th International Conference on Agents and Artificial Intelligence. Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, 538-546, 2025. Porto, Portugal: SCITEPRESS - Science and Technology Publications, 2025, pp. 538–546. DOI: [10.5220/0013158400003890](https://doi.org/10.5220/0013158400003890). URL: <https://hal.science/hal-04993041> (cit. on p. 27).
- [26] R. Bawden, M. Bénard, E. Villemonte de La Clergerie, J. Cornejo Cárcamo, N. Dahan, M. Delorme, M. Huguin, N. Kübler, P. Lerner, A. Mestivier, J. Minder, J.-F. Nominé, Z. Peng, L. Romary, P. Tsolakis, L. Zhu and F. Yvon. ‘MaTOS: Machine Translation for Open Science’. In: *Proceedings of Machine Translation Summit XX: Volume 2*. 20th Machine Translation Summit. Vol. Volume 2, Project Presentation Papers. Geneva, Switzerland, 24th June 2025. URL: <https://hal.science/hal-05228687> (cit. on p. 29).
- [27] R. Bawden and B. Sagot. ‘RoCS-MT v2 at WMT 2025: Robust Challenge Set for Machine Translation’. In: *Proceedings of the Tenth Conference on Machine Translation. WMT 2025 - Tenth Conference on Machine Translation*. Suzhou, China, 2025, pp. 834–849. URL: <https://hal.science/hal-05344725> (cit. on pp. 29, 31).
- [28] F. Cafiero, L. Ing, S. Gabay and T. Clérice. ‘“I am too old for this style!” A stylometric benchmark of age effect on authorship attribution’. In: *CHR 2025 - 6th Conference on Computational Humanities Research*. Luxembourg, Luxembourg, 2025. URL: <https://inria.hal.science/hal-05169869> (cit. on p. 44).
- [29] T. Clérice and A. Pinche. ‘Wauchier, Is That You? A multi-manuscript authorship analysis of Saint Lambert’s life’. In: *Anthology of Computers and the Humanities. CHR 2025 - 6th Conference on Computational Humanities Research*. Vol. 3. Luxembourg, Luxembourg, 2025, pp. 149–165. DOI: [10.63744/QsBV0XYj8wRC](https://doi.org/10.63744/QsBV0XYj8wRC). URL: <https://inria.hal.science/hal-05154115> (cit. on p. 44).
- [30] F. Dekmak, C. Khairallah and W. Antoun. ‘TutorMind at BEA 2025 Shared Task: Leveraging Fine-Tuned LLMs and Data Augmentation for Mistake Identification’. In: *BEA 2025 - 20th Workshop on Innovative Use of NLP for Building Educational Applications*. Vol. Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025). Vienna, Austria: Association for Computational Linguistics, 31st July 2025, pp. 1203–1211. DOI: [10.18653/v1/2025.bea-1.96](https://doi.org/10.18653/v1/2025.bea-1.96). URL: <https://hal.science/hal-05448906>.
- [31] R. Dent, P. Ortiz Suarez, T. Clérice and B. Sagot. ‘Identifying Rare Languages in Common Crawl Data is a Needles-in-a-Haystack Problem’. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 1460–1473. DOI: [10.18653/v1/2025.findings-emnlp.77](https://doi.org/10.18653/v1/2025.findings-emnlp.77). URL: <https://inria.hal.science/hal-05361348> (cit. on p. 42).
- [32] M. Futral, C. Schmid, B. Sagot and R. Bawden. ‘Towards Zero-Shot Multimodal Machine Translation’. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Findings of the Association for Computational Linguistics: NAACL 2025. Albuquerque, New Mexico, United States, 2025, pp. 761–778. URL: <https://hal.science/hal-04736377> (cit. on p. 33).

- [33] M. Futral, A. Zebaze, P. O. Suarez, J. Abadji, R. Lacroix, C. Schmid, R. Bawden and B. Sagot. ‘mOSCAR: A Large-scale Multilingual and Multimodal Document-level Corpus’. In: ACL 2025 - Findings of the Association for Computational Linguistics. Vienna, Austria, July 2025, pp. 3461–3494. DOI: [10.18653/v1/2025.findings-acl.180](https://doi.org/10.18653/v1/2025.findings-acl.180). URL: <https://hal.science/hal-04629451> (cit. on pp. 27, 33).
- [34] A. Garí Soler, M. Labeau and C. Clavel. ‘Potentially Problematic Word Usages and How to Detect Them: A Survey’. In: 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025). Suzhou, China, 8th Nov. 2025. URL: <https://hal.science/hal-05337734> (cit. on p. 38).
- [35] A. Garí Soler, M. Labeau and C. Clavel. ‘Toward the Automatic Detection of Word Meaning Negotiation Indicators in Conversation’. In: EMNLP 2025 - 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China, 4th Nov. 2025. URL: <https://hal.science/hal-05337731> (cit. on p. 38).
- [36] Z. M. Hüsünbeyi, D. Seddah and T. Scheffler. ‘Integrating Semantic Representations in a Cross-Modal Approach to Fact-Checking’. In: *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation, MAD’25*. MAD’25: 4th ACM International Workshop on Multimedia AI against Disinformation. Chicago USA, United States: ACM, 30th June 2025, pp. 17–27. DOI: [10.1145/3733567.3735567](https://doi.org/10.1145/3733567.3735567). URL: <https://hal.science/hal-05449205> (cit. on p. 37).
- [37] J. Janes, R. Bawden, T. Clérice, R. Dent, L. Ing, O. Nédey and B. Sagot. ‘Encoding language diversity in TEI: a description of regional and non-standard languages in multilingual diachronic corpora’. In: TEI Conference 2025. Cracovie, Poland, 15th Sept. 2025. URL: <https://hal.science/hal-05249807> (cit. on p. 42).
- [38] B. Kiessling. ‘Version 5 of the Kraken ATR Engine for the Humanities’. In: ICDAR2025 - 19th International Conference on Document Analysis and Recognition. Wuhan, China, 16th Sept. 2025. URL: <https://inria.hal.science/hal-05144723> (cit. on p. 43).
- [39] C. de Kock, A. Riabi, Z. Talat, M. S. Schlichtkrull, P. Madhyastha and E. Hovy. ‘IYKYK: Using language models to decode extremist cryptotexts’. In: EACL 2026 - 19th Conference of the European Chapter of the Association for Computational Linguistics. Rabat, Morocco, 5th June 2025. URL: <https://hal.science/hal-05480721>.
- [40] T. Kocmi, E. Artemova, E. Avramidis, R. Bawden, O. Bojar, K. Dranch, A. Dvorkovich, S. Dukanov, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, M. Karpinska, P. Koehn, H. Lakoungna, J. Lundin, C. Monz, K. Murray, M. Nagata, S. Perrella, L. Proietti, M. Popel, M. Popović, P. Riley, M. Shmatova, S. Steingrímsson, L. Yankovskaya and V. Zouhar. ‘Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets’. In: *Proceedings of the Tenth Conference on Machine Translation*. WMT25 - Tenth Conference on Machine Translation. Suzhou, China: Association for Computational Linguistics, 5th Nov. 2025, pp. 355–413. DOI: [10.18653/v1/2025.wmt-1.22](https://doi.org/10.18653/v1/2025.wmt-1.22). URL: <https://inria.hal.science/hal-05406602> (cit. on p. 29).
- [41] M. Lafon, Y. Karmim, J. Silva-Rodríguez, P. Couairon, C. Rambour, R. Fournier-Sniehotta, I. B. Ayed, J. Dolz and N. Thome. ‘ViLU: Learning vision-language uncertainties for failure prediction’. In: ICCV 2025. International Conference on Computer Vision. Honolulu, United States, Oct. 2025. URL: <https://hal.science/hal-05383056>.
- [42] M. Marmonier, R. Bawden and B. Sagot. ‘Explicit Learning and the LLM in Machine Translation’. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 12th Mar. 2025, pp. 31360–31410. DOI: [10.18653/v1/2025.emnlp-main.1599](https://doi.org/10.18653/v1/2025.emnlp-main.1599). URL: <https://inria.hal.science/hal-04991098> (cit. on p. 32).
- [43] M. Marmonier, B. Sagot and R. Bawden. ‘A French Version of the OLDI Seed Corpus’. In: *Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*, WMT 2025 - Tenth Conference on Machine Translation. Suzhou, China: Association for Computational Linguistics, 2025, pp. 1048–1060. DOI: [10.18653/v1/2025.wmt-1.80](https://doi.org/10.18653/v1/2025.wmt-1.80). URL: <https://hal.science/hal-05375157> (cit. on pp. 33, 41).

- [44] A. Ngo, N. Rollet, C. Pelachaud and C. Clavel. ‘“Mm, Wat?” Detecting Other-initiated Repair Requests in Dialogue’. In: *EMNLP 2025 - Conference on Empirical Methods in Natural Language Processing*. Suzhou, China, 4th Nov. 2025. URL: <https://hal.science/hal-05335583> (cit. on p. 38).
- [45] C. Nouri, J.-P. Cointet and C. Clavel. ‘Graphically Speaking: Unmasking Abuse in Social Media with Conversation Insights’. In: *63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*. Vienna, Austria, 27th July 2025. URL: <https://sciencespo.hal.science/hal-05165879> (cit. on p. 37).
- [46] Z. Peng, R. Bawden and F. Yvon. ‘Self-Retrieval from Distant Contexts for Document-Level Machine Translation’. In: *Proceedings of the Tenth Conference on Machine Translation. WMT 2025 - Conference on Machine Translation*. Suzhou, China, Nov. 2025. URL: <https://hal.science/hal-05353046>.
- [47] A. Riabi, V. Moulleron, M. Mahamdi, W. Antoun and D. Seddah. ‘Beyond Dataset Creation: Critical View of Annotation Variation and Bias Probing of a Dataset for Online Radical Content Detection’. In: *COLING 2025 - 31st International Conference on Computational Linguistics*. Abu Dhabi, United Arab Emirates, 19th Jan. 2025. URL: <https://hal.science/hal-04867863> (cit. on p. 36).
- [48] N. R. Robinson, C. Bizon Monroc, R. Dent, S. Watson, K. Murray, R. Dabre, A. Coy and H. Lent. ‘Findings of the First Shared Task for Creole Language Machine Translation at WMT25’. In: *WMT 2025 - Tenth Conference on Machine Translation*. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 520–531. DOI: [10.18653/v1/2025.wmt-1.28](https://doi.org/10.18653/v1/2025.wmt-1.28). URL: <https://inria.hal.science/hal-05361426> (cit. on p. 42).
- [49] S. C. Wan, M. Labeau and C. Clavel. ‘EmoDynamiX: Emotional Support Dialogue Strategy Prediction by Modelling MiXed Emotions and Discourse Dynamics’. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics. Albuquerque, United States: Association for Computational Linguistics, 29th Apr. 2025, pp. 1678–1695. DOI: [10.18653/v1/2025.naacl-1.ong.81](https://doi.org/10.18653/v1/2025.naacl-1.ong.81). URL: <https://hal.science/hal-05466118>.
- [50] A. R. Zebaze, B. Sagot and R. Bawden. ‘Compositional Translation: A Novel LLM-based Approach for Low-resource Machine Translation’. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 6th Mar. 2025, pp. 22328–22357. DOI: [10.18653/v1/2025.findings-emnlp.1216](https://doi.org/10.18653/v1/2025.findings-emnlp.1216). URL: <https://hal.science/hal-05009363> (cit. on p. 32).
- [51] A. R. Zebaze, B. Sagot and R. Bawden. ‘In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation’. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Findings of the Association for Computational Linguistics: NAACL 2025. Findings of the Association for Computational Linguistics: NAACL 2025. Albuquerque, United States: Association for Computational Linguistics, 2025, pp. 1222–1252. DOI: [10.18653/v1/2025.findings-naacl.68](https://doi.org/10.18653/v1/2025.findings-naacl.68). URL: <https://inria.hal.science/hal-04669351> (cit. on p. 32).
- [52] A. R. Zebaze, B. Sagot and R. Bawden. ‘TopXGen: Topic-Diverse Parallel Data Generation for Low-Resource Machine Translation’. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China, 2025, pp. 22358–22381. DOI: [10.18653/v1/2025.findings-emnlp.1217](https://doi.org/10.18653/v1/2025.findings-emnlp.1217). URL: <https://inria.hal.science/hal-05318504> (cit. on p. 32).

National peer-reviewed Conferences

- [53] A. Barkar, M. Chollet, M. Labeau, B. Biancardi and C. Clavel. ‘Décoder le pouvoir de persuasion dans les concours d’éloquence : une étude sur la capacité des modèles de langues à évaluer la prise de parole en public’. In: *Actes de l’atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*. 20e Conférence en Recherche d’Information et Applications (CORIA) 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 27ème Rencontre des

- Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) Les 18e Rencontres Jeunes Chercheurs en RI (RJCRI). Marseille, France: ATALA & ARIA, 2025, pp. 77–90. URL: <https://inria.hal.science/hal-05329786> (cit. on p. 27).
- [54] O. Nédey. ‘La traduction automatique dialectale: état de l’art et étude préliminaire sur le continuum dialectal de l’occitan’. In: *Actes des 18e Rencontres Jeunes Chercheurs en RI (RJCRI) et 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*. 20e Conférence en Recherche d’Information et Applications (CORIA) 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) Les 18e Rencontres Jeunes Chercheurs en RI (RJCRI). Marseille, France: ATALA & ARIA, 2025, pp. 190–238. URL: <https://inria.hal.science/hal-05330659> (cit. on p. 33).
- [55] B. Sagot, S. Ouni, S. Bigeard, L. Ing, R. Dent, J. Janes, T. Clérice, R. Bawden, E. Vincent, O. Nédey, M. Yaich, P. Tsolakis, V. Colotte and M. Sadeghi. ‘COLaF: Corpus and Tools for Languages of France and Varieties of French’. In: *Actes de la session industrielle de CORIA-TALN 2025*. 20e Conférence en Recherche d’Information et Applications (CORIA) 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) Les 18e Rencontres Jeunes Chercheurs en RI (RJCRI). Marseille, France: ATALA & ARIA, 2025, pp. 33–47. URL: <https://inria.hal.science/hal-05330342> (cit. on p. 40).
- [56] O. Silem, M. Fleig, P. Blache, H. Oufaida and L. Becerra-Bonache. ‘Une Approche Linguistique pour l’Évaluation des Caractéristiques du Langage Parlé dans les Modèles Conversationnels’. In: *Actes de l’atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*. 20e Conférence en Recherche d’Information et Applications (CORIA) 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) Les 18e Rencontres Jeunes Chercheurs en RI (RJCRI). Marseille, France: ATALA & ARIA, 2025, pp. 277–290. URL: <https://inria.hal.science/hal-05329785>.

Conferences without proceedings

- [57] T. Clérice. ‘Corpus Liberatum Linguæ Graecæ (CLLG): Libérer les textes en grec ancien’. In: *Antiqui.TXTes - Sciences des textes anciens*. Lyon, France, 2nd Sept. 2025. URL: <https://hal.science/hal-05245305> (cit. on p. 44).
- [58] T. Clérice. ‘Des données de projet aux données de projets : exemples de réutilisations de données’. In: *DHNNord2025. Valoriser les données de recherche en humanités numériques : enjeux, pratiques, perspectives*. Tourcoing, France, 2025. URL: <https://hal.science/hal-05421660>.
- [59] T. Clérice. ‘Distributed Text Services for Digital Classics’. In: *Digital Classicist London Seminar 2025*. Londres, United Kingdom, 11th July 2025. URL: <https://inria.hal.science/hal-05160204> (cit. on p. 44).
- [60] T. Clérice. ‘Présentation des Distributed Text Services’. In: *Antiqui.TXTes - Sciences des textes anciens*. Lyon, France, 2nd Sept. 2025. URL: <https://hal.science/hal-05245313> (cit. on p. 44).
- [61] R. Dent, T. Clérice, P. Ortiz Suarez and B. Sagot. ‘Français Tirailleur and Tâÿ Bôi: Institution-Driven Pidginization?’ In: *Twelfth Creolistics Workshop 2025*. Aarhus, Denmark, 3rd Dec. 2025. URL: <https://inria.hal.science/hal-05449221> (cit. on p. 42).
- [62] R. Dent, T. Clérice, P. Ortiz Suarez and B. Sagot. ‘Towards a Network-based Approach to French-Creole Diachrony’. In: *SPCL 2025 - Society for Pidgin and Creole Linguistics Summer Meeting*. Mona, Jamaica, 25th June 2025. URL: <https://inria.hal.science/hal-05376418> (cit. on p. 42).

- [63] S. Gabay, T. Hodel, R. Sluijter, É. Paupe, J.-C. Rebetez, D. Rabouin, V. Giovannangeli, W. Boente, É. Bascoul, M. Philip, M.-L. Massot, V. Ventresque, S. Crespi, P. Jacsont, Y. Jauregui, L. Chappuis, E. Solé, E. Zimmermann, M. Humeau, M. Lamrayah, J. Faldiola and A. Chagué. ‘Transcribing Western modern manuscripts (1500-2020): An economical, ecological and secured approach’. In: DH 2025 - Digital Humanities conference. Lisbon, Portugal, 14th July 2025. URL: <https://hal.science/hal-05063299> (cit. on p. 43).
- [64] J. Janes, S. Bénéière, B. Sagot and T. Clérice. ‘A TEI-based Layout Annotation System for a Deeper Automatic Encoding of Documents’. In: TEI 2025 - New Territories - TEI Conference and Members’ Meeting 2025. Krakow, Poland, Apr. 2025. URL: <https://inria.hal.science/hal-05232691> (cit. on p. 42).
- [65] B. Kiessling. ‘Large Multilingual ATR Models and Humanities Practice’. In: 2025 Workshop SCOOP - Source Codes of the Past. Princeton, NJ, United States, 10th June 2025. URL: <https://inria.hal.science/hal-05150070> (cit. on p. 43).
- [66] O. Nédey, J. Janes, T. Clérice, R. Bawden and B. Sagot. ‘Retour d’expérience sur la création d’un corpus en occitan à partir de textes produits par des utilisateurs’. In: Journée d’études AFIA-ATALA-AFCP « Technologies linguistiques pour les langues peu dotées » (TLLPD). Paris, France, 12th Dec. 2025. URL: <https://inria.hal.science/hal-05407710> (cit. on p. 41).

Scientific book chapters

- [67] Y. Xu, Y. Prado, R. Severson, S. Lovato and J. Cassell. ‘Growing Up with Artificial Intelligence: Implications for Child Development’. In: *Children and Screens: A Handbook on Digital Media and the Development, Health, and Well-being of Children and Adolescents*. Springer Nature Switzerland, 6th Dec. 2025, pp. 611–617. DOI: [10.1007/978-3-031-69362-5_83](https://doi.org/10.1007/978-3-031-69362-5_83). URL: <https://inria.hal.science/hal-04986368>.

Doctoral dissertations and habilitation theses

- [68] L. Nishimwe. ‘Robust Neural Machine Translation of User-Generated Content’. Sorbonne Université, 18th June 2025. URL: <https://theses.hal.science/tel-05448644> (cit. on p. 31).
- [69] A. Riabi. ‘Small is Beautiful : addressing resource scarcity, language variation, and transfer challenges for automatic detection of Harmful language’. Sorbonne Université, 18th Mar. 2025. URL: <https://theses.hal.science/tel-05123132> (cit. on p. 36).

Reports & preprints

- [70] S. Arias, M. Bergmann, F. Campillo, M.-A. Enard, C. Fabre, F. Garcia, B. Guedj, E. Jeannot, G. Neglia, D. Peurichard, D. Racoceanu, B. Sagot and G. Tworkowski. *Reflections on the Use of Generative AI for Research Professions*. Inria, 9th July 2025. URL: <https://inria.hal.science/hal-05188001>.
- [71] S. Arias, M. Bergmann, F. Campillo, M.-A. Enard, C. Fabre, F. Garcia, B. Guedj, E. Jeannot, G. Neglia, D. Peurichard, D. Racoceanu, B. Sagot and G. Tworkowski. *Réflexions sur l’usage de l’IA générative pour les métiers de la recherche*. Inria, 2025, pp. 1–10. URL: <https://inria.hal.science/hal-05187992>.
- [72] P. Chambon, B. Roziere, B. Sagot and G. Synnaeve. *BigO(Bench) – Can LLMs Generate Code with Controlled Time and Space Complexity?* 21st Mar. 2025. URL: <https://hal.science/hal-05000733> (cit. on p. 27).
- [73] T. Clérice, S. Gabay, M. Vlachou-Efstathiou, A. Pinche and B. Sagot. *CoMMA, a Large-scale Corpus of Multilingual Medieval Archives*. 9th Dec. 2025. URL: <https://inria.hal.science/hal-05299220> (cit. on p. 44).
- [74] R. Dent, P. Ortiz Suarez, T. Clérice and B. Sagot. *KréyoLID: From Language Identification Towards Language Mining*. 9th Mar. 2025. URL: <https://inria.hal.science/hal-04986402> (cit. on p. 42).

- [75] M. Gille Levenson, O. Brisville-Fertin, M. Castillo Lluch, M. D. Yáñez, T. Faye, S. Gabay, L. Ing, M. Labrousse, M. López Izquierdo, C. Pascual-Argente, A. Pinche, I. Salvo García, M. B. Villar Díaz, F. Le Guen and K. Desnoues. *Manuel d'annotation du projet e-CaM: Étiquetage lexico-grammatical du castillan médiéval*. ENS de Lyon; Agorantic FR 3621; CIHAM, 1st Dec. 2025. URL: <https://hal.science/hal-05390100> (cit. on p. 45).
- [76] N. Godey, W. Antoun, R. Touchent, R. Bawden, É. de la Clergerie, B. Sagot and D. Seddah. *Gaperon: A Peppered English-French Generative Language Model Suite*. 29th Oct. 2025. URL: <https://hal.science/hal-05410121> (cit. on p. 26).
- [77] N. Godey, A. Devoto, Y. Zhao, S. Scardapane, P. Minervini, É. de la Clergerie and B. Sagot. *Q-Filters: Leveraging QK Geometry for Efficient KV Cache Compression*. 4th Mar. 2025. URL: <https://hal.science/hal-05441473> (cit. on p. 26).
- [78] F. Gouzi, T. Tasovac, A. Baillot, S. Bénière, C. Delmazo and V. Garnett. *ATRIUM Peer Review Framework (Version 1)*. DARIAH ERIC, 30th Dec. 2025. DOI: [10.5281/zenodo.17787914](https://doi.org/10.5281/zenodo.17787914). URL: <https://shs.hal.science/halshs-05458734>.
- [79] J. Janes, S. Bénière, L. Ing and T. Clérice. *LADaS Annotation Guidelines*. 12th Sept. 2025. URL: <https://inria.hal.science/hal-05252327> (cit. on p. 43).
- [80] B. Kiessling. *Transcription Guidelines for Generalized Automatic Text Recognition*. 22nd Dec. 2025. URL: <https://hal.science/hal-05429033> (cit. on p. 43).
- [81] B. Kiessling, P. Stokes, L. Romary, T. Hodel, C. Kermorvant, S. Gabay, M. Gille Levenson, O. Brisville-Fertin, M. Vlachou-Efstathiou, M. Guénette, A. von Stockhausen, M. Verstraete, R. Chauhan, M. Bizais-Lillig, C. Vidal-Gorène, A. Kasparian, A. Tanelian, A. Ohanian, N. Lucas, A. Perrier and C. Salah. *Apprendre à Lire aux Machines*. Ed. by A. Chagué, T. Clérice and A. Pinche. 11th July 2025. URL: <https://hal.science/hal-05163931> (cit. on p. 43).
- [82] T. Kocmi, E. Avramidis, R. Bawden, O. Bojar, K. Dranch, A. Dvorkovich, S. Dukanov, N. Fedorova, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, M. Karpinska, P. Koehn, H. Lakoungna, J. Lundin, K. Murray, M. Nagata, S. Perrella, L. Proietti, M. Popel, M. Popović, P. Riley, M. Shmatova, S. Steingrímsson, L. Yankovskaya and V. Zouhar. *Preliminary Ranking of WMT25 General Machine Translation Systems*. 2025. URL: <https://inria.hal.science/hal-05406662> (cit. on p. 29).
- [83] B. Mohapatra, T. Charlot, G. Duca, M. Palan, L. Romary and J. Cassell. *Frame of Reference: Addressing the Challenges of Common Ground Representation in Situational Dialogs*. 14th Jan. 2026. DOI: [10.48550/arXiv.2601.09365](https://doi.org/10.48550/arXiv.2601.09365). URL: <https://hal.science/hal-05479462> (cit. on p. 46).
- [84] O. Nédey, J. Janès, R. Bawden, T. Clérice and B. Sagot. *ForumOccitania: a Corpus of User-Generated Content for Multiple Occitan Varieties*. 12th Dec. 2025. URL: <https://inria.hal.science/hal-05413035> (cit. on p. 41).
- [85] L. Nishimwe, B. Sagot and R. Bawden. *When the Gold Standard isn't Necessarily Standard: Challenges of Evaluating the Translation of User-Generated Content*. 19th Dec. 2025. URL: <https://hal.science/hal-05426685> (cit. on p. 31).
- [86] Z. Peng, R. Bawden and F. Yvon. *Model Cards for the MaTOS Project*. Projet ANR MaTOS, 15th Aug. 2025. URL: <https://inria.hal.science/hal-04803089> (cit. on p. 30).
- [87] C. R. M. A. Santagiustina and P. Ramaciotti Morales. *Representation, Political Discontent and e-Petitions: How economic ideology and anti-elitism drive online petitioning in the EU*. 30th Nov. 2025. URL: <https://hal.science/hal-05389645> (cit. on p. 28).
- [88] H. Scheithauer, G. Romero-García, L. Romary and T. Clérice. *Les formats d'encodage de la notation musicale à l'épreuve d'un objectif de transcription manuscrite automatique*. 2025. URL: <https://inria.hal.science/hal-05478071>.
- [89] A. G. Soler, J. Myrendal, C. Clavel and S. Larsson. *The NeWMe Corpus: A gold standard corpus for the study of Word Meaning Negotiation*. 23rd Apr. 2025. DOI: [10.21203/rs.3.rs-5975927/v1](https://doi.org/10.21203/rs.3.rs-5975927/v1). URL: <https://inria.hal.science/hal-05423801> (cit. on p. 38).

- [90] A. R. Zebaze, R. Bawden and B. Sagot. *LLM Reasoning for Machine Translation: Synthetic Data Generation over Thinking Tokens*. 13th Oct. 2025. URL: <https://inria.hal.science/hal-05318507> (cit. on p. 32).
- [91] Y. Zuo, K. Gerdes, E. Villemonte de La Clergerie and B. Sagot. *Patent Representation Learning via Self-supervision*. 31st Oct. 2025. URL: <https://hal.science/hal-05333463> (cit. on p. 40).

Other scientific publications

- [92] E. Bobrov, L. Bracco, M. Dacos, N. Fressengeas, I. Hrynaszkiewicz, A. Iarkaeva, A. Peršić, V. Proudman, L. Romary and R. Sabo. *The Principles of Open Science Monitoring*. 21st July 2025. DOI: [10.5281/zenodo.15807481](https://doi.org/10.5281/zenodo.15807481). URL: <https://hal.science/hal-05162786>.
- [93] T. Clérice. *Referee report. For: The artificial intelligence cooperative: READ-COOP, Transkribus, and the benefits of shared community infrastructure for automated text recognition [version 1]*. 2025. DOI: [10.21956/openreseurope.20281.r50056](https://doi.org/10.21956/openreseurope.20281.r50056). URL: <https://hal.science/hal-04957346>.
- [94] T. Clérice, R. Bawden, D. A. Smith and A. Pinche. ‘PaRAMHTRS (Philology And Resolution of Abbreviations in Manuscripts obtained by HTR at Scale)’. In: *Journées du Datalab 2026*. Paris, France, 20th Jan. 2026. URL: <https://inria.hal.science/hal-05453316> (cit. on p. 31).
- [95] L. Ing and M. Gille Levenson. ‘Variant modelization (in multilingual context)’. In: *Workshop 2025 Seeing the Difference: Visualizing Textual Variation*. Leiden, Netherlands, 14th Apr. 2025. URL: <https://hal.science/hal-05173463> (cit. on p. 45).
- [96] O. Nédey, T. Clérice, R. Bawden and B. Sagot. ‘Machine Translation for Low-Resource Dialectal Variants’. In: *Journée de l’école doctorale EDITE*. Paris, France, 11th Mar. 2025. URL: <https://inria.hal.science/hal-05366306> (cit. on p. 33).

Educational activities

- [97] T. Clérice. ‘Building a cross-lingual dataset from medieval manuscript text recognition Challenges and outcomes of CATMuS’. *École thématique*. United Kingdom, 20th Jan. 2025. URL: <https://hal.science/hal-05150349> (cit. on p. 43).

12.3 Cited publications

- [98] E. M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell. ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?’ In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. DOI: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922). URL: <https://doi.org/10.1145/3442188.3445922> (cit. on p. 14).
- [99] G. Castillo-López, A. Riabi and D. Seddah. ‘Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection’. In: *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1–13. DOI: [10.18653/v1/2023.vardial-1.1](https://doi.org/10.18653/v1/2023.vardial-1.1). URL: <https://inria.hal.science/hal-04243810> (cit. on p. 34).
- [100] R. Dent, J. Janes, T. Clérice, P. Ortiz Suarez and B. Sagot. ‘Molyé: A Corpus-based Approach to Language Contact in Colonial France’. In: *NLP4DH 2024 - 4th International Conference on Natural Language Processing for Digital Humanities*. 8 main pages and 3 pages of references. Miami, United States, Nov. 2024. URL: <https://hal.science/hal-04736370> (cit. on p. 42).
- [101] S. Desrochers, C. Paradis and V. M. Weaver. ‘A Validation of DRAM RAPL Power Measurements’. In: *Proceedings of the Second International Symposium on Memory Systems*. MEMSYS ’16. Alexandria, VA, USA: Association for Computing Machinery, 2016, pp. 455–470. DOI: [10.1145/2989081.2989088](https://doi.org/10.1145/2989081.2989088). URL: <https://doi.org/10.1145/2989081.2989088> (cit. on p. 15).

- [102] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova. ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://aclanthology.org/N19-1423> (cit. on p. 10).
- [103] P.-A. Duquenne, H. Schwenk and B. Sagot. ‘SONAR: Sentence-Level Multimodal and Language-Agnostic Representations’. working paper or preprint. Oct. 2023. URL: <https://inria.hal.science/hal-04264028> (cit. on p. 31).
- [104] D. Elliott, S. Frank, K. Sima’an and L. Specia. ‘Multi30K: Multilingual English-German Image Descriptions’. In: *Proceedings of the 5th Workshop on Vision and Language*. Ed. by A. Belz, E. Erdem, K. Mikolajczyk and K. Pastra. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 70–74. DOI: [10.18653/v1/W16-3210](https://doi.org/10.18653/v1/W16-3210). URL: <https://aclanthology.org/W16-3210/> (cit. on p. 33).
- [105] F. Feng, Y. Yang, D. Cer, N. Arivazhagan and W. Wang. ‘Language-agnostic BERT Sentence Embedding’. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by S. Muresan, P. Nakov and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 878–891. DOI: [10.18653/v1/2022.acl-long.62](https://doi.org/10.18653/v1/2022.acl-long.62). URL: <https://aclanthology.org/2022.acl-long.62/> (cit. on p. 30).
- [106] J. Foster. ‘“cba to check the spelling”: Investigating Parser Performance on Discussion Forum Posts’. In: *NAACL*. Los Angeles, California, 2010 (cit. on p. 10).
- [107] M. Futral, C. Schmid, I. Laptev, B. Sagot and R. Bawden. ‘Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. for Computational Linguistics. Vol. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, July 2023, pp. 5394–5413. DOI: [10.18653/v1/2023.acl-long.295](https://doi.org/10.18653/v1/2023.acl-long.295). URL: <https://inria.hal.science/hal-03977982> (cit. on pp. 27, 33).
- [108] Y. Guo, G. Shang, M. Vazirgiannis and C. Clavel. ‘The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text’. In: *NAACL 2024 Findings - Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Accepted to NAACL 2024 Findings. Mexico City, Mexico, June 2024. URL: <https://hal.science/hal-04593399> (cit. on p. 28).
- [109] C. Lothritz, B. Lebichot, K. Allix, S. Ezzini, T. Bissyandé, J. Klein, A. Boytsov, C. Lefebvre and A. Goujon. ‘Evaluating the Impact of Text De-Identification on Downstream NLP Tasks’. In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Ed. by T. Alumäe and M. Fishel. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 10–16. URL: <https://aclanthology.org/2023.nodalida-1.2> (cit. on p. 35).
- [110] Y. Lv and C. Zhai. ‘When documents are very long, BM25 fails!’ In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’11. Beijing, China: Association for Computing Machinery, 2011, pp. 1103–1104. DOI: [10.1145/2009916.2010070](https://doi.org/10.1145/2009916.2010070). URL: <https://doi.org/10.1145/2009916.2010070> (cit. on p. 30).
- [111] J. Maillard, C. Gao, E. Kalbassi, K. R. Sadagopan, V. Goswami, P. Koehn, A. Fan and F. Guzman. ‘Small Data, Big Impact: Leveraging Minimal Data for Effective Machine Translation’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 2740–2756. DOI: [10.18653/v1/2023.acl-long.154](https://doi.org/10.18653/v1/2023.acl-long.154). URL: <https://aclanthology.org/2023.acl-long.154/> (cit. on p. 41).

- [112] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean. ‘Distributed Representations of Words and Phrases and their Compositionality’. In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. 2013, pp. 3111–3119. URL: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality> (cit. on p. 10).
- [113] S. Montariol, A. Riabi and D. Seddah. ‘Multilingual Auxiliary Tasks Training: Bridging the Gap between Languages for Zero-Shot Transfer of Hate Speech Detection Models’. In: *Findings of ACL 2022*. Accepted to Findings of ACL-IJCNLP 2022. Online, France, Nov. 2022. URL: <https://inria.hal.science/hal-03840070> (cit. on p. 34).
- [114] A. Ngo, D. Heylen, N. Rollet, C. Pelachaud and C. Clavel. ‘Exploration of Human Repair Initiation in Task-oriented Dialogue : A Linguistic Feature-based Approach’. In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Kyoto, Japan, Sept. 2024, pp. 603–609. URL: <https://hal.science/hal-04745323> (cit. on p. 38).
- [115] L. Nishimwe, B. Sagot and R. Bawden. ‘Making Sentence Embeddings Robust to User-Generated Content’. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti and N. Xue. Accepted at LREC-COLING 2024. Torino, Italy, 2024, pp. 10984–10998. URL: <https://hal.science/hal-04520909> (cit. on p. 31).
- [116] R. Nordquist. *Linguistic Variation*. ThoughtCo. 2019. URL: <https://www.thoughtco.com/what-is-linguistic-variation-1691242> (cit. on p. 10).
- [117] P. J. Ortiz Suárez, B. Sagot and L. Romary. ‘Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures’. In: *Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Ed. by P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lungen and C. Iliadi. Cardiff, United Kingdom: Leibniz-Institut für Deutsche Sprache, July 2019. DOI: [10.14618/IDS-PUB-9021](https://doi.org/10.14618/IDS-PUB-9021). URL: <https://hal.inria.fr/hal-02148693> (cit. on p. 33).
- [118] A. Pellicani, G. Pio, D. Redavid and M. Ceci. ‘SAIRUS: Spatially-aware identification of risky users in social networks’. In: *Information Fusion* 92 (2023), pp. 435–449. DOI: <https://doi.org/10.1016/j.inffus.2022.11.029>. URL: <https://www.sciencedirect.com/science/article/pii/S1566253522002457> (cit. on p. 35).
- [119] B. Plank. ‘The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation’. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10671–10682. DOI: [10.18653/v1/2022.emnlp-main.731](https://doi.org/10.18653/v1/2022.emnlp-main.731). URL: <https://aclanthology.org/2022.emnlp-main.731/> (cit. on p. 35).
- [120] A. Riabi, M. Mahamdi, V. Moulleron and D. Seddah. ‘Cloaked Classifiers: Pseudonymization Strategies on Sensitive Classification Tasks’. In: *Proceedings of the fifth Workshop on Privacy in Natural Language Processing*. Bangkok, Thailand, Aug. 2024. URL: <https://inria.hal.science/hal-04624789> (cit. on p. 35).
- [121] M. Sanguinetti, C. Bosco, L. Cassidy, Ö. Çetinoğlu, A. T. Cignarella, T. Lynn, I. Rehbein, J. Ruppenhofer, D. Seddah and A. Zeldes. ‘Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations’. In: *Language Resources and Evaluation* 57.2 (Feb. 2022), pp. 493–544. DOI: [10.1007/s10579-022-09581-9](https://doi.org/10.1007/s10579-022-09581-9). URL: <https://hal.science/hal-04629571> (cit. on p. 10).
- [122] R. Schwartz, J. Dodge, N. A. Smith and O. Etzioni. ‘Green AI’. In: *Commun. ACM* 63.12 (Nov. 2020), pp. 54–63. DOI: [10.1145/3381831](https://doi.org/10.1145/3381831). URL: <https://doi.org/10.1145/3381831> (cit. on p. 14).
- [123] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. Li, Y. Wu and D. Guo. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. 2024. URL: <https://arxiv.org/abs/2402.03300> (cit. on p. 27).

- [124] E. Strubell, A. Ganesh and A. McCallum. ‘Energy and Policy Considerations for Deep Learning in NLP’. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3645–3650. doi: [10.18653/v1/P19-1355](https://doi.org/10.18653/v1/P19-1355). URL: <https://aclanthology.org/P19-1355> (cit. on pp. 14, 16).
- [125] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith and Y. Choi. ‘Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics’. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 9275–9293. doi: [10.18653/v1/2020.emnlp-main.746](https://doi.org/10.18653/v1/2020.emnlp-main.746). URL: <https://aclanthology.org/2020.emnlp-main.746/> (cit. on p. 34).
- [126] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk and J. Wang. *No Language Left Behind: Scaling Human-Centered Machine Translation*. 2022. arXiv: [2207.04672](https://arxiv.org/abs/2207.04672) [cs.CL]. URL: <https://arxiv.org/abs/2207.04672> (cit. on p. 33).
- [127] R. Touchent, L. Romary and E. De La Clergerie. ‘CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé’. In: *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*. Ed. by C. Servan and A. Vilnat. Paris, France: ATALA, June 2023, pp. 323–334. URL: <https://hal.science/hal-04130187> (cit. on pp. 13, 39).
- [128] R. Touchent, L. Romary and E. De La Clergerie. ‘CamemBERT-bio: Leveraging Continual Pre-training for Cost-Effective Models on French Biomedical Data’. In: *LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. Torino, Italy, May 2024. URL: <https://hal.science/hal-04528508> (cit. on pp. 13, 39).
- [129] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts and C. Raffel. ‘ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models’. In: *Transactions of the Association for Computational Linguistics* 10 (2022). Ed. by B. Roark and A. Nenkova, pp. 291–306. doi: [10.1162/tacl_a_00461](https://doi.org/10.1162/tacl_a_00461). URL: <https://aclanthology.org/2022.tacl-1.17/> (cit. on p. 31).