Département
D4: Natural Language Processing & Knowledge Discovery

Équipe SMarT

Statistical Machine Translation Speech Modelization and Text

Loria

Laboratoire lorrain de recherche en informatique et ses applications

Rapport d'activité 2025

cnrs UNIVERSITÉ DE LORRAINE
En partenariat avec
Inria CentraleSupélec

Contents

Team SMarT

Keywords

Applications for under-resourced languages (Machine translation, Automatic speech recognition,...), Frugal AI, Code-switching, Enhancement of esophageal speech, Pathological speech recognition, Opinion mining, Automatic music generation, Tracking fakes news.

1 Team members, visitors, external collaborators

Faculty Members

- Kamel Smaïli [Team leader, UL, Professor, HDR]
- Joseph di Martino [UL, Assistant Professor (Emeritus]
- Sahbi Sidhom [UL, Assistant Professor, HDR]

PhD Students

- Ouahab Hocini [Université Lorraine, PhD Student, , until 2026]
- Yassine Toughrai [UL, PhD Student, , until Dec. 2026]
- Margaux Adloff [UL, PhD Student, , until Dec. 2028]
- Amina Laggoun [UL, PhD Student, , until Dec. 2028]

Administrative Assistant

- Anne-Marie Messaoudi [CNRS]

External Collaborators

- Youness Moukafih [UIR, Université Internationale Rabat, HDR]
- Ouassim Karrakchou [UIR, Université Internationale Rabat, HDR]
- Sihem Zakaria [ESI, Université d'Alger, HDR]
- Tahar Kechadi [UCD, University College of Dublin, HDR]
- Mikolaj Leszczuk [Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
- Karim Bouzoubaa [Université Mohammed V, Rabat, HDR]
- Chiraz Latiri [University of Tunis – Tunisia, HDR]

2 Overall objectives

The research objective of SMarT is to innovate in various aspects of Natural Language Processing (NLP) by developing novel methods and resources. These methods primarily focus on leveraging deep learning techniques to transform textual or speech signals into other forms. This transformation has broad applications including machine translation, automatic speech recognition, enhancing voice clarity, generating musical accompaniments for vocal melodies, question-answering systems, and fine-tuning sentiment analysis. SMarT specializes in addressing challenges associated with under-resourced languages such as Arabic dialects or pathological voices.

At SMarT, our present focus lies in tackling the challenge of fake news detection, a pressing issue given the proliferation of rumors across social media platforms. This endeavor is central to the ongoing TRADEF project spearheaded by our research group[11, 38]. Additionally, we aim to extend our efforts to address a broader concern: Deepfakes [40], a term derived from the fusion of Deep learning and fake news, both key research areas within SMarT. Deepfakes represent a more sophisticated and go beyond fake news, leveraging mediums beyond text, such as manipulated videos and images featuring prominent individuals, with potentially detrimental effects on public perception. This work began in the Phd of Ouahab Hocini.

The challenge of training a model becomes exponentially more complex as the number of parameters increases. Take GPT-3, the largest language model to date, for example, it utilizes a staggering 175 billion parameters and was trained on a corpus of 500 billion words. GPT-3 serves as a neural language model, enabling tasks like natural language generation, classification, question-answering, and summarization. Training GPT-3 on a single GPU would have taken 355 years, but thankfully, it was developed using a large number of processors in parallel. This underscores the limitations of classical methods in addressing certain problems. This research topic aims to address the challenges of processing low-resource languages by leveraging unstructured and non-standardized data from social media platforms such as Twitter, Facebook, and TikTok. Despite being noisy and inconsistent, these data sources provide valuable insights into the real usage of Maghrebi Arabic dialects in natural contexts. To overcome the lack of annotated data and pre-trained models, the approach relies on self-supervised learning techniques that enable models to learn from large amounts of unlabeled data. Models such as BERT can be adapted to capture contextual representations of dialectal Arabic from raw social media text. However, modern NLP models require substantial computational resources and large datasets, which are often unavailable for under-resourced languages. This constraint motivates the adoption of frugal AI, which focuses on designing efficient models with limited computational and energy resources. The research also highlights several scientific challenges, including orthographic variability, code-switching, and noise in social media data, which complicate preprocessing and dialect identification. Additionally, data scarcity and bias affect the representativeness of corpora and the generalization ability of models. The design of lightweight models through techniques such as quantization and knowledge distillation introduces a trade-off between efficiency and performance. Finally, the project explores multi-criteria [23] decision-making methods and tools like ShrinkBench to optimize model efficiency, reduce energy consumption, and ensure reliable and explainable NLP systems for dialectal Arabic.

3 Research program

3.1 Processing and providing resources to Arabic dialects

One of the most important part of the research work in SM$_{a}$rT, since one decade, is to provide new models and resources to Arabic dialects [15, 30, 33, 25, 37, 29, 13, 32, 27, 14, 28, 34, 12, 26, 24, 36, 35]. In fact, Arabic language comprises thirty modern varieties[1], including its standard form, Modern Standard Arabic (MSA) which is the official form used in the newspapers and in the formal communications. The other Arabic language varieties, referred as dialects, come from historical interactions between classical Arabic and languages of the regional cultures and from the linguistic influence due to colonization. They are used in the Arab world in informal conversational context and in the daily communication. Due to their varieties and because they are not official languages, they suffer from a lack of resources. Therefore, they constitute a real challenge, not only because we have to collect more and more data to develop deep learning models for processing them, but also because they are under-resourced languages and we have to consider them as such.

One of the results of 2025 was the development of knowledge distillation to create TinyDziriBERT, a compact BERT-based model for the low-resource Algerian dialect, using DziriBERT as the teacher model. TinyDziriBERT is approximately seven times smaller than DziriBERT (70.8 MB vs. 498 MB) but retains competitive performance. On Twifil sentiment analysis, the distilled model achieves 78.39%

[1] https://iso639-3.sil.org/code/ara

accuracy, closely approaching DziriBERT's 79.86% and far exceeding the non-distilled baseline (71.08%). For emotion detection, TinyDziriBERT reaches 66.04% accuracy versus 70.27% for the teacher. In dialect identification, the best distilled model attains an F1 score of 76.12%, nearly matching DziriBERT's 77.34%. Notably, TinyDziriBERT outperforms larger multi-dialectal models like mBERT and AraBERT on dialect-specific tasks, achieving 79.21% accuracy on the BOUTEF dataset compared to 72.33% for mBERT. Overall, TinyDziriBERT enables efficient deployment of Algerian dialect NLP on resource-constrained devices.

3.2 Detecting fake news

This research concerns the TRADEF project, an ASTRID/ANR-funded initiative devoted to tracking and detecting fake news and deepfakes in Arab social networks. It builds on the BOUTEF corpus, a multilingual and multi-dialect resource containing posts in Algerian and Tunisian dialects, MSA, French, English, and mixed scripts, together with comments and metadata. The study highlights several manipulation strategies used in disinformation, including altered spellings, sensational language, fake authority, and media imitation, especially in narratives targeting Emmanuel Macron. To address this problem, we propose a multimodal workflow combining fake-news classification, sentiment analysis of comments, named-entity extraction, semantic comparability, and audio–subtitle consistency checking for videos. On the modeling side, the research concerns the use and fusion of Arabic language models such as AraBERT, DarijaBERT, MarBERT, DziriBERT, mBERT, and XLM-R, with the proposed consistency-aware fusion achieving the best results [31]. A second major contribution is ABDUL [39], a family of language models trained only on formal corpora but adapted to dialects through phonemic normalization, showing strong transfer to low-resource dialect tasks.

Another activity concerning the misinformation strated in the thesis (ENACT) of Margaux Adloff focuses on the study of disinformation in Large Language Models (LLMs), using genocide-related narratives as a critical case study. The work aims to clarify the concept of disinformation in the context of AI, where traditional notions such as intention, truth, and harm become difficult to define due to the non-deterministic and non-intentional nature of LLMs. To support this investigation, several corpora were constructed, covering three genocides—the Armenian Genocide, the Holocaust, and the Rwandan Genocide—each divided into factual historical texts and denialist discourse. A detailed corpus analysis, particularly on Holocaust-related data, was conducted using statistical methods and linguistic markers to distinguish between factual and distorted narratives.

In parallel, the research includes a strong theoretical component, drawing on linguistic and philosophical frameworks, such as the theory of information by Barwise and Perry, to better understand the relationship between information, meaning, and knowledge. This allows for a deeper conceptualization of disinformation in AI-generated content. The study also engages with contemporary debates on LLMs, including their relationship to truth, bias, and user perception, as well as philosophical questions concerning machine intentionality inspired by Searle.

From a technical perspective, the research explores the fine-tuning of LLMs on curated corpora to evaluate their ability to detect denialist strategies, comparing trained and untrained models. Overall, this work constitutes a multidisciplinary research activity combining corpus construction, NLP, and philosophy to better understand and mitigate disinformation in AI systems.

3.3 Pathological voices and advanced voice anonymization

One of the researches of SM$_{a^r}$T is to enhance pathological voices. For that, we investigated several ways to attend this objective. One of the tracks was the detection of the fundamental frequency of speech based on time-frequency decomposition of the cepstrum using wavelets [16]. The results obtained, using two international databases, are quite interesting and far exceeded the results obtained by state of the art works. One of the most important research areas concerning this axis is related to the recognition of both laryngeal and alaryngeal speech. We improved the recognition of esophageal speech using a hybrid system based on statistical voice conversion. We also improved the phonetic recognition accuracy through the projection, by a statistical voice conversion technique, of esophageal cepstral vectors into a laryngeal space. One of the research axes of SM$_{a^r}$T concerns the vocal conversion, in other words, the transformation of a laryngeal voice A into a laryngeal voice B and by vocal correction the transformation of a pathological voice into a laryngeal voice [17, 18, 19, 21, 22, 20]. One of the main ideas explored in these

works is to extract the excitation and phase related to the converted cepstra directly from the training space. By doing so, in the case of voice conversion, the identity of the target speaker is preserved and in the case of voice correction, the converted voice tends to a laryngeal voice. A Sequence-to-Sequence model and deep learning architectures have been used to perform the transformation of the esophageal voice into a laryngeal one.

Another activity in this research domain focuses on preserving vocal privacy in response to the growing use of speech technologies. The primary objective was to design a modular anonymization pipeline capable of transforming a speaker's voice into a protected identity. The system aims to decouple biometric traits from the linguistic message, ensuring a balance between user privacy and the naturalness of the synthesized speech. The technical architecture relies on a disentanglement strategy to separate speech components. The process begins with Wav2Vec 2.0, which extracts phonetic content independently of the speaker. Simultaneously, an ECAPA-TDNN encoder captures the original identity, which is then mapped to a new "pseudo-speaker" target. The anonymization is strictly a one-way transformation, ensuring that the original speaker's biometric identity cannot be recovered or reversed from the output. To ensure a fluid output, FastSpeech 2 is used for generative prosody, while a HiFi-GAN vocoder reconstructs the final high-quality audio signal. The evaluation of the framework confirmed its effectiveness in protecting against automatic speaker recognition while maintaining high intelligibility. Future work will explore expanding the system to support multiple languages and integrating emotional control to allow for more versatile and expressive anonymized voices.

3.4 Research at the Intersection of Information Sciences and AI

Recent work outlines a coherent scientific trajectory at the intersection of information and communication sciences and digital sciences. This research is structured around three main axes. The first axis focuses on semantic engineering for augmented intelligence, formalized in a monograph Dynamics of Knowledge in Information-Communication Artifacts [8] and applied to intelligent platforms for digital commons and heritage resources [5]. The second axis addresses intelligent information retrieval systems and user-centered mediation, integrating connected objects and AI models with a particular emphasis on explainability and usage, as developed in contributions to STICEF and the CIAREI conference [4]. The third axis explores inferential mechanisms and the production of actionable knowledge, notably through knowledge graphs and decision-support systems applied to fields such as medical imaging, including editorial coordination of a special issue on AI for medical imaging. These contributions, largely first-authored, are complemented by epistemological and ethical reflections on artificial intelligence [10]. Together, they support the development of a research program oriented toward augmented, explainable information retrieval systems capable of generating operational knowledge.

4 Application domains

The research activities of SM$_a{}^r$T has several applications. The research group operates at the intersection of Natural Language Processing (NLP), artificial intelligence, and digital humanities, with a particular focus on Arabic dialects and emerging AI paradigms. One of its primary application domains is the processing and analysis of Arabic dialectal data, especially from the Maghreb region. This includes tasks such as dialect identification, sentiment analysis, named entity recognition, and the handling of code-switching phenomena (e.g., Arabizi), using both monolingual and multilingual language models. In parallel, the group explores the detection of fake news and misinformation, particularly in multilingual and dialectal contexts. By leveraging Large Language Models (LLMs) and smaller, task-specific models, the research investigates how generated rationales and explainability techniques can improve the reliability of automated verification systems. This work is especially relevant in the context of social media, where information is often noisy, unstructured, and rapidly evolving. Another important application area is frugal AI, which focuses on designing efficient and sustainable machine learning models. The group develops and evaluates techniques such as model compression, knowledge distillation, and pruning to reduce computational costs and energy consumption, while maintaining strong performance. This is particularly critical for low-resource languages and environments with limited infrastructure. Additionally,

the research extends to the analysis and production of music, with a specific interest in Arabic and Andalusian musical traditions. This includes the use of AI for music generation, style modeling, and the study of musical structures, bridging the gap between technology and cultural heritage. Finally, the group engages in broader investigations Large Language Models (LLMs), including their prompting strategies, reasoning capabilities, and adaptation to low-resource and dialectal settings. By combining these diverse domains, the research aims to develop robust, explainable, and resource-efficient AI systems tailored to real-world linguistic and cultural challenges.

5 New software, platforms, open data

5.1 Open data

- BOUTEF, a multilingual fake news corpus, contains 57095 genuine and fake narrative entries in Arabic dialects, Modern Standard Arabic (MSA), French, English, and code-switching formats. Each post is associated to 18 metadata.

6 New results

6.1 International research visitors

6.1.1 Visits of international scientists

6.1.2 Visits to international teams

- Kamel Smaïli: Multiple visits to UIR (Université Internationale de Rabat), engaging in discussions and collaborate with Dr Ouassim Karrakchou and Youness Moukafih on innovative methods aimed at addressing the challenges posed by Arabic dialects especially in the domain of frugal AI. The primary goal is to develop effective applications capable of handling these linguistic variations efficiently. This leads to a funded phd thesis that started on December 2025

- Visiting USTHB and ESI in Algeria, two of the leading institutions in computer science, to deliver talks and engage with students. These visits helped attract strong candidates to the lab, including Amina Laggoun, whom I supervised during her Master's thesis and who is now pursuing her PhD at Loria.

6.2 European initiatives

6.2.1 MUTASK - Multimodal Translation and Adaptation of Scientific Knowledge for Global Accessibility (2025-2028)

SM$_a$rT proposed and participated to a project developing European AI-driven methods to translate and adapt scholarly content, bridging linguistic gaps and ensuring vital research reaches diverse audiences. It confronts three main hurdles: language barriers limiting engagement with scientific findings, overly dense content for lay audiences, and a lack of effective workflows to deliver adapted materials. To tackle these challenges, the consortium integrates machine translation, automated summarization, semantic indexing, and video-based storytelling into one streamlined workflow. A core goal is to reshape complex academic documents into dynamically narrated videos, detailed yet understandable summaries, and localized translations accessible at multiple expertise levels. Users can pause videos to request clarifications, ensuring a truly adaptive learning experience with a human touch centered on media education and strategic communication. By joining specialists from Poland, France, and Switzerland, the project delivers practical tools that fit current publishing systems while upholding Open Science standards. The project directly addresses the CHIST-ERA 2025 "Science in Your Own Language" call by developing multimodal translation and knowledge adaptation solutions for scholarly content. It employs AI-based translation to convert research articles across multiple European languages, handling specialized jargon and document-level complexity. Outputs are designed for integration into open data repositories like CLARIN and EOSC, ensuring interoperable access to multilingual content. Through

iterative testing with students, journalists, and nonprofits, the project fosters more inclusive, democratic, and trusted science by ensuring knowledge in any language or format remains comprehensible and widely disseminated.

6.3 National initiatives

6.3.1 TRADEF - Astrid (2023-2026)

The 4th generation warfare (4GW) is known as information warfare involving populations that are not necessarily military. It is waged by national or transnational groups following ideologies based on cultural convictions, religion, economic or political interests, with the aim of sowing chaos in a targeted part of the world. In 1989, the authors of an article on fourth-generation warfare, some of whom are military, explained that fourth-generation warfare would be widespread and difficult to define in the decades to come.

With the emergence of social networks, the battlefield whose contours were blurred has found a place for 4GW. Indeed, one of the penetration points of 4GW is the massive use of social networks to manipulate opinions, the aim being to prepare opinion in one part of the world to accept a state of affairs and make it humanly acceptable and politically correct.

Like fourth-generation warfare, cognitive warfare aims to blur the mechanisms by which politics, economics, religion, etc. are understood. The consequence of this action is to destabilise and reduce the adversary. This cognitive war therefore targets the brain of what is supposed to be the enemy. In the end, the new battlefield moves to the brain of the adversary or, more precisely, to the subconscious of the adversary's population.

The aim of this war is to alter reality by, among other things, often inundating the adversary's population with false information, rumours and fabricated or modified videos.

What's more, the proliferation of social bots today means that disinformation can be generated automatically on social networks. According to some sources, during the 2016 US elections, 19% of the total volume of tweets generated were thanks to these automatic bots.

In TRADEF, a ANR ASTRID project for which we are the leader (collaboration with the Laboratoire d'Informatique d'Avignon), we are looking at a few areas of disinformation: fake news, deep fake and potentially harmful information. The idea is to detect very quickly in social networks, the birth of a fake in its textual, audio or video form and its propagation through the networks. The idea is to detect the birth of a fake and track it over time. At any given moment, this potential rumour is analysed and assigned a confidence rating, and tracked across social networks in the language of reference as well as in other languages. As the suspect information evolves over time, its score will change according to the data with which it is confronted. The information to be tested is matched with audio or video data that can confirm or refute the credibility of the information. Videos that can be used as sources to denounce a fake can themselves be deepfakes. This leads us to be vigilant when examining these videos by developing robust deepfake detection methods. Finally, this project introduces a dimension of explicability into the results.

Given the experience of the participating teams in deep learning and automatic processing of the standard Arabic language and its dialects, we propose to track and identify fakes and potentially harmful information in Arabic social networks, which will give rise to other scientific challenges such as the processing of code-switching, the variability of Arabic dialects, the identification of named entities in the speech continuum, the development of neural methods for languages with few resources and the explicability of the results obtained.

7 Dissemination

7.1 Promoting scientific activities

- Kamel Smaïli was appointed in 2025 to the Scientific Council of the Scientific and Technical Research Center for the Development of the Arabic Language, Algiers, Algeria.

7.1.1 Scientific events: organisation

General chair, scientific chair

Member of the organizing committees

7.1.2 Scientific events: selection

Chair of conference program committees

Member of the conference program committees We are member of several conferences in the domain: Interspeech, ICASSP, ICALP, JEP, TALN and others.

Reviewer The Team members were reviewers for Interspeech, LREC, ACL, Interspeech and ICASSP since several years.

Reviewer - reviewing activities The members of SMarT regularly provide their expertise in reviewing journal articles, conference papers, ANR projects, and CIFRE proposals.

7.1.3 Invited talks

- Kamel Smaïli was invited as a speaker invited to the FIRST SEMINAR on Large language models LLMs: and give a talk : "From Probabilities to Power: The Rise of LLMs ", April 7, 2025
- Kamel Smaïli has been invited to give a talk as an invited speaker at Ecole Nationale d'Informatique organisé par le laboratoire LCMS, "The Future of Low-Resource NLP: Exploring Powerful Neural Network Architectures", June 30, 2025,
- Kamel Smaïli gives a tal at University of Mohammed V, Rabat, Morroco, May 22, 2024. " The future of low-resource NLP: Exploring powerful neural network architectures architectures, the case of SeamlessM4T"
- Kamel Smaïli has been invited by the university Mohammed V (ENSIAS) as an invited speaker to give a talk in the context 5th Edition of the Doctoral Days of Arabic Language Engineering, "Tackling the Complexities of Arabic Dialects", 18-19 July 2025.

7.1.4 Scientific expertise

In addition to ANR and CIFRE, the team was involved in 2022 in the evaluation of projects of NeuroInsights, where the objective of these kind of projects to propose new projects allowing to understand how machine learning can be used to understand, detect and find new ways to diagnose, treat and manage neurological conditions ranging from relatively minor disorders to serious chronic and rare diseases.

7.1.5 Research administration

- Kamel Smaïli oversees the management of the SMarT research group.
- Kamel Smaïli heads a national project (ASTRID): TRADEF.

7.2 Teaching - Supervision - Juries

Kamel Smaïli participated to a dozen of Phd and masters committees theses last year in: France, Morocco and Tunisia.

7.2.1 Teaching

- Kamel Smaïli contributes to the master NLP of university of Lorraine, Master MIAGE (France and Morocco) and Master II.
- Sahbi Sidhom contributes to the Master "veille stratégique et organisation des connaissances" (Strategic Information Management; Knowledge organization seminar)

7.2.2 Supervision

Kamel Smaïli supervised the phd of the following students:

1. Fadi Al-Ghawnmeh, defended September 25, 2025: arab and AI: Maqām Music Generation through Machine Translation and Motion Capture, Universite of Oslo
2. Ouahab Hocini (2022-2026): Fake News Detection in Arabic Dialects with Consistency-Aware LLM Merging Techniques
3. Yassine Toughrai(2023-2026): Tracking fake news in Arabic social networks.
4. Margaux Adloff (2025-2028): From Information to Disinformation: Analyzing LLM Behavior on Sensitive Historical Corpora
5. Amina Laggoun (2025-2028): Explainability ad Frugal AI for under-resourced languages

7.3 Popularization

7.3.1 Internal or external responsibilities

- Sahbi Sidhom founded ISKO-Maghreb[2] chapter and is president of this chapter (Société savante ISKO-Maghreb - Chapter (Tunisie, Algérie, Maroc, Libye et Liban)

8 Scientific production

8.1 Major publications

[1] A. Hocini and K. Smaïli. 'Boosting Fake News Detection in Arabic Dialects with Consistency-Aware LLM Merging Techniques'. In: 6th International Conference on Natural Language Processing and Computational Linguistics (NLPCL 2025). Toronto (CA), Canada, 27th Sept. 2025, pp. 25–33. DOI: 10.5121/csit.2025.151803. URL: https://hal.science/hal-05365859.

[2] Y. Toughrai, D. Langlois and K. Smaïli. 'Fake News Detection via Intermediate-Layer Emotional Representations'. In: Companion Proceedings of the ACM on Web Conference 2025. Sydney, Australia: ACM, 28th Apr. 2025, pp. 2680–2684. DOI: 10.1145/3701716.3717537. URL: https://hal.science/hal-04977819.

[3] Y. Toughrai, K. Smaïli and D. Langlois. 'ABDUL: a new Approach to Build language models for Dialects Using formal Language corpora only'. In: NAACL 2025 Workshop on Language Models for Underserved Communities. Albuquerque, United States, 4th Mar. 2025. URL: https://hal.science/hal-05004706.

8.2 Publications of the year

Invited conferences

[4] S. Sidhom. 'Interaction, Mediation and Intelligence: Towards the Integration of Communicating Objects into Intelligent Systems.' In: Congrès International CIAREI : Communication, intelligence artificielle, remédiation, éthique et inclusion. Strabourg, France, 29th Sept. 2025. URL: https://hal.univ-lorraine.fr/hal-05021978.

[2] www.isko-maghreb.org

International peer-reviewed conferences

[5] S. Sidhom. 'Towards an Intelligent Platform Serving Digital Commons: Semantic and Collaborative Enrichment of Heritage Resources.' In: HyperHeritage International Symposium (HIS). Vol. Vol. 1. Paris Campus Condorcet, France, 3rd Nov. 2025, pp.1–23. URL: https://hal.univ-lorraine.fr/hal-05392790.

Conferences without proceedings

[6] A. Laggoun, C. Zakaria and K. Smaïli. 'Knowledge Distillation for Efficient Algerian Dialect Processing: Training Compact BERT Models with DziriBERT'. In: 7th International Conference on Advances in Signal Processing and Artificial Intelligence. Innsbruck (Austria), Austria, 8th Apr. 2025. URL: https://hal.science/hal-04998510.

[7] Y. Toughrai, K. SmaÃ¯li and D. Langlois. 'Modeling North African Dialects from Standard Languages'. In: The Third Arabic Natural Language Processing Conference. Suzhou, China, 2nd Nov. 2025. URL: https://hal.science/hal-05348485.

Scientific books

[8] S. Sidhom. *Dynamique de la connaissance dans les artefacts information-communication: Cadres théoriques, perspectives Ã©pistÃ©mologiques et processus dâ€™intelligence*. Vol. Vol. I.D Â« Ã‰mergences, cheminements I.D et constructions de savoirs Â», SÃ©rie 2 : Â« Information, Documentation et construction de savoirs Â». L'Harmattan, 16th June 2025, p. 238. URL: https://hal.univ-lorraine.fr/hal-05121397.

[9] S. Sidhom, L. L. Jilani and J. Belhadj. *Organization of Knowledge and Advanced Technologies: Artificial Intelligence*. Vol. Vol. 1, Vol. 2, Vol. 3. International Multi-Conference OCTA on Organization of Knowledge and Advanced Technologies No. 4, 2025. 20th Nov. 2025, p. 640. URL: https://hal.univ-lorraine.fr/hal-05392711.

Reports & preprints

[10] M. Trestini, C.-A. Magot, M. Zeyringer, J.-F. Plateau, M. Frisch, P.-Y. Connan, F. Emprin, A. Promonet, J.-L. P. Bergey, P. Viallon, S. Sidhom, B. Coulibaly, M. Denami, H. Sabra, S. Mazziotti, A. Collin, N. Bannier, Ã. Rabin, A. Grandadam and D. Deas. *Results of the teacher-student survey on the uses and non-uses of Generative AI in France's eastern academic region*. GTnum IA2GE de la Direction du numérique Ã©ducatif du MENESR, 15th May 2025. URL: https://hal.science/hal-05062217.

8.3 Cited publications

[11] H. Abdelouahab and K. Smaïli. 'Detecting Fake News: Exploring Key Features in Multilingual Arabic Dialect Corpus'. In: *The 8th International Conference on Arabic Language Processing*. RABAT, Morocco, Apr. 2024. URL: https://hal.science/hal-04578312.

[12] K. Abidi, M. A. Menacer and K. Smaili. 'CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube'. In: *18th Annual Conference of the International Communication Association (Interspeech)*. Conference of the International Communication Association (Interspeech). Stockholm, Sweden, Aug. 2017. URL: https://hal.archives-ouvertes.fr/hal-01531591.

[13] K. Abidi and K. Smaïli. 'An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings'. In: *11th edition of the Language Resources and Evaluation Conference, LREC 2018*. Miyazaki, Japan, May 2018. URL: https://hal.archives-ouvertes.fr/hal-01718110.

[14] K. Abidi and K. Smaïli. 'An empirical study of the Algerian dialect of Social network'. In: *ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing*. Casablanca, Morocco, Dec. 2017. URL: https://hal.inria.fr/hal-01659997.

[15] M. Amine Menacer, K. Abidi, N. Othman and K. Smaïli. 'Sentiment analysis of videos in Algerian dialect'. In: *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*. Ed. by C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla and S. Schneider. Nancy, France: ATALA, 2020, pp. 296–304. URL: https://hal.archives-ouvertes.fr/hal-02784779.

[16] F. Bahja, J. Di Martino, E. H. Ibn Elhaj and D. Aboutajdine. 'A corroborative study on improving pitch determination by time–frequency cepstrum decomposition using wavelets'. In: *SpringerPlus* (2016). DOI: 10.1186/s40064-016-2162-0. URL: https://hal.inria.fr/hal-01312747.

[17] I. Ben Othmane, J. Di Martino and K. Ouni. 'A novel voice conversion approach using cascaded powerful cepstrum predictors with excitation and phase extracted from the target training space encoded as a KD-tree'. In: *International Journal of Speech Technology* (2019), pp. 1–13. DOI: 10.1007/s10772-019-09643-4. URL: https://hal.inria.fr/hal-02315052.

[18] I. Ben Othmane, J. Di Martino and K. Ouni. 'Enhancement of esophageal speech using statistical and neuromimetic voice conversion techniques'. In: *Journal of International Science and General Applications* 1.1 (2018), p. 10. URL: https://hal.inria.fr/hal-01724375.

[19] I. Ben Othmane, J. Di Martino and K. Ouni. 'Enhancement of esophageal speech obtained by a voice conversion technique using time dilated Fourier cepstra'. In: *International Journal of Speech Technology* 22.1 (2018), pp. 99–110. DOI: 10.1007/s10772-018-09579-1. URL: https://hal.inria.fr/hal-01954096.

[20] I. Ben Othmane, J. Di Martino and K. Ouni. 'Enhancement of esophageal speech using voice conversion techniques'. In: *International Conference on Natural Language, Signal and Speech Processing - ICNLSSP 2017*. Casablanca, Morocco, Dec. 2017. URL: https://hal.inria.fr/hal-01660580.

[21] I. Ben Othmane, J. Di Martino and K. Ouni. 'Improving the computational performance of standard GMM-based voice conversion systems used in real-time applications'. In: *ICECOCS'18 - 1st International Conference on Electronics, Control, Optimization and Computer Science*. Kenitra, Morocco, Dec. 2018. DOI: 10.1109/ICECOCS.2018.8610514. URL: https://hal.inria.fr/hal-01886099.

[22] I. B. Ben Othmane, J. Di Martino and K. Ouni. 'Vers la transformation de la parole oesophagienne en voix laryngée à l'aide de techniques de conversion vocale'. In: *7ème Journées de Phonétique Clinique - JPC 7*. Paris, France, June 2017. URL: https://hal.inria.fr/hal-01563783.

[23] K. Govindan. 'How Artificial Intelligence Drives Sustainable Frugal Innovation: A Multitheoretical Perspective'. In: *IEEE Transactions on Engineering Management* 71 (2024), pp. 638–655. DOI: 10.1109/TEM.2021.3116187.

[24] S. Harrat, K. Meftouh, M. Abbas, W.-K. Hidouci and K. Smaïli. 'An Algerian dialect: Study and Resources'. In: *International journal of advanced computer science and applications (IJACSA)* 7.3 (2016), pp. 384–396. DOI: 10.14569/IJACSA.2016.070353. URL: https://hal.archives-ouvertes.fr/hal-01297415.

[25] S. Harrat, K. Meftouh, K. Abidi and K. Smaïli. 'Automatic identification methods on a corpus of twenty five fine-grained Arabic dialects'. In: *Arabic Language Processing: From Theory to Practice7th International Conference, ICALP 2019, Nancy, France, October 16–17, 2019, Proceedings*. Vol. Communications in Computer and Information Science book series (CCIS, volume 1108). Oct. 2019. DOI: 10.1007/978-3-030-32959-4_6. URL: https://hal.archives-ouvertes.fr/hal-02314245.

[26] S. Harrat, K. Meftouh and K. Smaïli. 'Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid'. In: *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*. Budapest, Hungary, Apr. 2017. URL: https://hal.archives-ouvertes.fr/hal-01557405.

[27] S. Harrat, K. Meftouh and K. Smaïli. 'Machine translation for Arabic dialects (survey)'. In: *Information Processing and Management* 56.2 (2017), pp. 262–273. DOI: 10.1016/j.ipm.2017.08.003. URL: https://hal.archives-ouvertes.fr/hal-01581038.

[28] S. Harrat, K. Meftouh and K. Smaïli. 'Maghrebi Arabic dialect processing: an overview'. In: *ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing*. ISGA. Casablanca, Morocco, Dec. 2017. URL: https://hal.inria.fr/hal-01660001.

[29] S. Harrat, K. Meftouh and K. Smaïli. 'Maghrebi Arabic dialect processing: an overview'. In: *Journal of International Science and General Applications* 1 (Mar. 2018). URL: https://hal.archives-ouvertes.fr/hal-01873779.

[30] S. Harrat, K. Meftouh and K. Smaïli. 'Script Independent Morphological Segmentation for Arabic Maghrebi Dialects: An Application to Machine Translation'. In: *Computación y sistemas* 23.3 (2019), pp. 979–989. DOI: 10.13053/cys-23-3-3267. URL: https://hal.archives-ouvertes.fr/hal-02274533.

[31] A. Hocini and K. Smaïli. 'Boosting Fake News Detection in Arabic Dialects with Consistency-Aware LLM Merging Techniques'. In: *6th International Conference on Natural Language Processing and Computational Linguistics (NLPCL 2025)*. Toronto (CA), Canada, Sept. 2025, pp. 25–33. DOI: 10.5121/csit.2025.151803. URL: https://hal.science/hal-05365859.

[32] K. Meftouh, S. Harrat and K. Smaïli. 'PADIC: extension and new experiments'. In: *7th International Conference on Advanced Technologies ICAT*. Antalya, Turkey, Apr. 2018. URL: https://hal.archives-ouvertes.fr/hal-01718858.

[33] K. Meftouh, K. Abidi, S. Harrat and K. Smaïli. 'The SMarT Classifier for Arabic Fine-Grained Dialect Identification'. In: *The Fourth Arabic Natural Language Processing Workshop co-located with ACL*. Florence, Italy, Aug. 2019. URL: https://hal.archives-ouvertes.fr/hal-02166384.

[34] M. A. Menacer, O. Mella, D. Fohr, D. Jouvet, D. Langlois and K. Smaïli. 'Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect'. In: *ACLing 2017 - 3rd International Conference on Arabic Computational Linguistics*. Dubai, United Arab Emirates, Nov. 2017, pp. 1–8. URL: https://hal.archives-ouvertes.fr/hal-01583842.

[35] Y. Moukafih, A. Ghanem, K. Abidi, N. Sbihi, M. Ghogho and K. Smaïli. 'SimSCL: A Simple fully-Supervised Contrastive Learning Framework for Text Representation'. In: *AJCAI 2021 - 34th Australasian Joint Conference on Artificial Intelligence*. Sydney, Australia, Feb. 2022. URL: https://hal.science/hal-03367972.

[36] Y. Moukafih, N. Sbihi, M. Ghogho and K. Smaïli. 'SuperConText: Supervised Contrastive Learning Framework for Textual representations'. In: *IEEE Access* (2023). DOI: 10.1109/ACCESS.2023.3241490. URL: https://hal.science/hal-03964804.

[37] K. Smaïli, ed. *Arabic Language Processing: From Theory to Practice*. Arabic Language Processing: From Theory to Practice 7th International Conference, ICALP 2019, Nancy, France, October 16–17, 2019, Proceedings. Loria - University Lorraine. Nancy, France: Springer, Oct. 2019. URL: https://hal.archives-ouvertes.fr/hal-03044112.

[38] K. Smaïli, A. Hamza-Jamann, L. David and A. Djegdjiga. 'BOUTEF: Bolstering Our Understanding Through an Elaborated Fake News Corpus'. In: *The 8th International Conference on Arabic Language Processing*. RABAT, Morocco, Apr. 2024. URL: https://hal.science/hal-04578297.

[39] Y. Toughrai, K. Smaïli and D. Langlois. 'ABDUL: a new Approach to Build language models for Dialects Using formal Language corpora only'. In: *NAACL 2025 Workshop on Language Models for Underserved Communities*. ACL. Albuquerque, United States, Apr. 2025. URL: https://hal.science/hal-05004706.

[40] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang and N. Yu. 'Multi-Attentional Deepfake Detection'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 2185–2194.