

AfIA

Association française
pour l'Intelligence Artificielle

Dossier N° 3

***P**anorama **F**rançais de la **R**echerche en **T**echnologies du **L**angage **H**umain*

Collège TLH

Septembre 2026 N° 3

AfIA
Association française
pour l'Intelligence Artificielle

SOMMAIRE
DU DOSSIER

Édito 3
ALTAE : Approches Linguistiques Théoriques, Appliquées et Expérimentales : langues et cultures connectée 4
AMIAD/Pôle Recherche : Groupe langage 12
MaIAGE/BIBLIOME : Acquisition et Formalisation de Connaissances à partir de Textes . . . 15
CEA-List/LASTI : Laboratoire Analyse Sémantique Texte Image 20
GREYC/DIAMA∩T : Document, Interactions, Apprentissage Multimodal, Accessibilité, tal ∩ vision, Texte 24
Inalco/ERTIM: Équipe de Recherche Textes, Informatique, Multilinguisme 28
IRISA/ARCHIMÉDIA 32
IRIT/MELODI : MEthodes et ingénierie des Langues, des Ontologies et du DIscours 34
LabHC : Laboratoire Hubert Curien 38
LaBRI : Laboratoire Bordelais de Recherche en Informatique. 41
LaTTICe : Langues, Textes, Traitements Informatiques, Cognition 45
LIA : Laboratoire Informatique d'Avignon 48
LIASD/GLAÇON : Groupe, Langage, Apprentissage, Corpus, évaluatiON 55
LIG/GETALP : Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole 58
LIS/TALEP : Traitement automatique du langage écrit et parlé 65
LISN/STL : Sciences et Technologies des Langues 69
LIUM/LST : Language and Speech Technologies. 75
LLF : Laboratoire de Linguistique Formelle 81
LORIA/Sémagramme : Analyse sémantique du langage naturel 85
LS2N/TALN : Traitement Automatique du Langage Naturel. 90
Orange Research 94
TETIS/GeoTexIA : GÉOinformations, données TEXtuelles et Intelligence Artificielle 100



AfIA
Association française
pour l'Intelligence Artificielle

Dossier réalisé par

Gaël DIAS
GREYC UMR 6072
Université de Caen Normandie
gael.dias@unicaen.fr

Corinne FREDOUILLE
LIA UPR 4128
Avignon Université
corinne.fredouille@univ-avignon.fr

Vincent P. MARTIN
LORIA UMR 7503
INRIA / Université de Lorraine
vincent.p.martin@inria.fr

AfIA
Association française
pour l'Intelligence Artificielle

Édito

Ce troisième dossier (éditions précédentes en 2020 et 2021) vise à compléter la liste des équipes de recherche académiques et industrielles françaises menant des travaux à l'intersection du traitement automatique des langues (TAL), de la recherche d'information (RI), de la communication parlée (CP) et de l'intelligence artificielle (IA), domaine communément appelé technologies du langage humain (TLH). En particulier, les TLH proposent des méthodes permettant une communication humain-machine naturelle, pouvant s'étendre à une interaction humain-humain médiée. Ainsi, les TLH permettent d'analyser, d'interpréter et de produire des actes du langage écrit, parlé ou signé, mais aussi d'interagir avec des données langagières.

Suite à ce nouvel appel à participation communiqué sur les listes de diffusion françaises des domaines de recherche des TLH, nous avons reçu 11 nouvelles contributions, totalisant ainsi 21 institutions, dont 20 laboratoires académiques et 1 entreprise, répartis sur 12 villes plus Paris et sa région (Figure 1). En plus des nouvelles contributions, une partie des équipes présentes dans les deux premiers bulletins (disponibles en ligne) ont également actualisé leurs participations. Ce bulletin pourra donc être croisé avec les deux précédentes éditions pour une vision plus complète des activités concernant le TLH en France.

La diversité des recherches présentées ainsi que la qualité et la quantité des contributions reçues démontrent à la fois une dynamique importante des TLH en France mais aussi un savoir-faire et des compétences reconnus à l'international. Notamment, il est très intéressant de remarquer la pluralité des approches scientifiques suivies, ce qui ne fait que renforcer une particularité nationale propice au foisonnement des idées.

Ce dossier ne se veut pas exhaustif mais a le mérite de rendre compte assez fidèlement du large spectre des thématiques abordées en TAL. Au-delà des grands thèmes de recherche contemporains sur les grands modèles de langues, la transcription automatique de la parole, la traduction automatique, ou les langues peu dotées qui sont toujours d'actualité, le panomara permet également de mettre au jour de nouveaux thèmes de recherche plus originaux. En particulier, ce nouveau bulletin met l'accent sur la forensic (deep fakes, confiance, biométrie), la fiabilité (évaluation, traçabilité, interprétabilité, souveraineté), l'éthique (biais, anonymisation, inclusion et accessibilité), la frugalité des modèles de langue, la multimodalité (texte, image, vidéo, audio, graphes de connaissances, etc.) et les approches centrées sur l'humain. Ces thèmes de recherche trouvent des applications aussi diverses que les sciences du vivant et la santé, l'écologie et l'agriculture, la sécurité ou les humanités numériques.



Fig. 1 : Cartographie des TLH en France.

Nous tenons à remercier particulièrement tous les contributeurs et contributrices de ce bulletin qui ont pris de leur temps et de leur énergie pour promouvoir leur discipline et informer la communauté de leurs recherches actuelles, ainsi que les membres du comité de pilotage du collège TLH pour leur soutien dans cette initiative.

Nous espérons que vous trouverez autant de plaisir à lire ce troisième dossier que nous en avons pris à sa réalisation. Bonne lecture.

Gaël Dias, Corinne Fredouille, Vincent P. Martin

AfIA
Association française
pour l'Intelligence Artificielle

ALTAE : Approches Linguistiques Théoriques, Appliquées et Expérimentales : langues et cultures connectée

ALTAE URP 3967
Université Paris Cité
https://altae.u-pariscite.fr

Nicolas BALLIER
nicolas.ballier@u-paris.fr

Maria ZIMINA-POIROT
maria.zimina-poirot@u-paris.fr

Présentation générale

ALTAE (Approches Linguistiques Théoriques, Appliquées et Expérimentales : langues et cultures connectées) est une Unité de Recherche Propre (URP 3967) d'Université Paris Cité, membre du consortium Empirical Foundations of Linguistics (EFL). Au sein d'Université Paris Cité, l'équipe ALTAE collabore régulièrement avec le Laboratoire de Linguistique Formelle (LLF), en particulier dans le cadre de la thématique 4 d'EFL (langage et IA). ALTAE, anciennement CLILLAC-ARP, réunit des experts développant des recherches en linguistique de l'anglais, de l'espagnol, de l'allemand et du français – de façon majoritaire – mais aussi en linguistique du vietnamien, du persan, du chinois, de l'espéranto, etc. ALTAE est fortement ancrée dans la linguistique de corpus et la linguistique appliquée. Elle est actuellement composée de quatre équipes :

- l'équipe APREM (Atelier Permanent de Réflexion Épistémologique et Méthodologique) aborde l'IA du point de vue de l'explicabilité et des biais de modèles ;
- l'équipe ET3C (Enjeux actuels de la Traduction et Communication des Cultures en Contact) envisage l'IA du point de vue de la qualité de la traduction automatique, de la post-édition, de la terminologie et de la phraséologie (spécialisées et transdisciplinaires), ainsi que du point de vue de la littératie numérique ;
- l'équipe 3DS utilise l'IA de manière exploratoire, entre autres pour les problématiques d'annotation ;
- l'équipe PAP (Phonétique, Acquisition & Psycholinguistique) utilise notamment l'IA pour la classification automatique des apprenants, la génération de stimuli contraints dans les expériences et la modélisation des données expérimentales.

Principaux thèmes de recherche

Littératie de la traduction automatique (TA)

Au sein de l'équipe ET3C, l'IA est abordée selon plusieurs axes de recherche : la qualité de la traduction automatique, la post-édition, la révision, la terminologie et la phraséologie (spécialisées et transdisciplinaires), ainsi que la littératie numérique. L'équipe a développé une expertise dans le domaine de l'analyse des erreurs en traduction, en particulier à travers des approches textuelles et des annotations, fondées sur des typologies d'erreurs. Ces travaux s'inscrivent dans des contextes théoriques, professionnels et pédagogiques, et visent à évaluer la qualité de la traduction humaine et automatique, ainsi que celle de la post-édition et de la révision. Elle a par ailleurs mis au point une nouvelle typologie adaptée à ces types de production mise en œuvre notamment dans le projet ANR MaTOS [13, 27].

Les membres d'ALTAE participent activement à l'évolution du référentiel de compétences des traducteurs professionnels à l'échelle européenne. Leur contribution s'inscrit notamment dans les travaux menés au sein du réseau EMT (*European Master's in Translation*), en lien avec les groupes de travail consacrés à l'évolution des compétences et à l'impact des technologies. Elle s'appuie également sur l'expertise reconnue des membres de l'équipe coresponsables des masters en traduction, en particulier pour l'intégration des outils numériques et de l'IA dans les programmes de formation [20, 18, 46].

L'équipe s'inscrit également dans plusieurs réseaux et partenariats institutionnels, notamment le réseau AFFUMT (l'Association française des formations universitaires aux métiers de la traduction), en lien avec les associations professionnelles, en par-

ticulier, la SFT (Société française des traducteurs) et la CNET (Chambre Nationale des Entreprises de Traduction), le réseau INSTB (*International Network of Simulated Translation Bureaus*), l'alliance européenne d'universités CIRCLE U.

Les membres d'ALTAE ont également publié plusieurs articles dans les médias écrits sur le rôle de l'intelligence artificielle en sciences humaines — notamment dans *Le Monde* ("Non, l'intelligence artificielle ne remplacera pas les traducteurs et les traductrices"), dans *Le Courrier de l'UNESCO* en 2022 et dans *Sciences humaines* en 2026 — ainsi que des vidéos présentant des interventions des enseignants-chercheurs responsables du master ILTS.

Adossée aux projets PAPTAN et MULTITAN-GML, notre participation régulière à des compétitions internationales telles que WMT (*World Machine Translation*), LREC et Interspeech pour des tâches de traduction nous a permis de renforcer nos compétences et de diffuser notre expertise dans une démarche de littératie numérique et de pédagogie de l'intelligence artificielle [12, 40, 41, 42].

L'équipe a également pris part à des projets Erasmus+ qui nourrissent une réflexion sur la formation des traducteurs à l'ère de l'IA, notamment le projet LT-LiDER (*Language and Translation : Literacy in Digital Environments and Resources*).

En collaboration avec le Centre d'Études de la Traduction, l'équipe développe un positionnement à la fois critique et analytique en traductologie face aux évolutions de l'intelligence artificielle et participe activement aux colloques spécialisés du domaine, notamment aux cycles de la conférence TRALOGY (*Translation and Technology*), dont la dernière édition a porté sur les apports et les défis de l'IA [23].

Intégration de l'IA dans la linguistique de corpus

Parallèlement à ces projets centrés sur la littératie numérique de la TA, l'équipe adapte des modèles de traduction neuronale pour plusieurs langues [36] et mène des expériences sur l'utilisation de l'IA pour l'analyse terminologique et la rédaction de définitions de termes [32], la simplification et la réécriture des textes [46, 44, 48], l'augmentation et l'enrichissement des données linguistiques, la gestion de projets en traduction et communication multilingue, tout en développant des workflows intégrant l'IA générative (IAG) pour ce type de tâches [44]. La mise en regard de l'approche fondée sur les corpus textuels avec celle reposant sur les outils d'IAG permet d'adopter une démarche expérimentale et critique de l'IAG, dans laquelle le corpus, en tant que réservoir de données linguistiques authentiques, garantit une analyse robuste et facilite l'évaluation des apports et des limites de l'IAG au sein du workflow outillé.

ALTAE est également engagée dans plusieurs partenariats industriels, à la fois dans les domaines de la pédagogie et de la recherche, en collaboration avec des acteurs technologiques tels que RWS, memoQ, Wordfast et ChapsVision [44].

Parallèlement à une recherche qui porte directement sur les représentations des modèles, nous explorons l'exploitation des GML (grands modèles de langage) dans nos analyses. Comme sans doute bien d'autres équipes, nous expérimentons avec les GML pour l'intégration de problématiques linguistiques dans le cadre de chaîne de traitements [29] notamment pour l'annotation [30] au service d'une linguistique très outillée. L'IA est alors pensée comme un outil susceptible d'opérationnaliser des cadres d'analyses, tels que la typologie d'erreurs de traduction développée au laboratoire [28], des problématiques de traductologie, de terminologie [32], ou de simplification textuelle [46, 44, 48] et plus généralement, de la linguistique appliquée [43]. Les GML sont alors utilisés comme comme outil de valorisation de ressources linguistiques développées au laboratoire, telle que la base de connaissance ARTES pour la terminologie ou le projet Outil de Collecte des Traductions des Apprenants en Vue de leur Exploration Scientifique (OCTAVES), application web permettant d'enrichir une base de données d'unités de traduction (phrases ou paragraphes) alignées. Nous développons une réflexion fondée sur notre expertise de linguistes en y associant des méthodes de la science des données et en s'appuyant sur les services de notre plateforme PNS-UP Pure Neural Server – plateforme de traduction automatique sur serveur université. Conçue pour des expérimentations en traduction menées en partenariat industriel avec SYSTRAN by ChapsVision, cette plateforme constitue un environnement numérique dédié

à l'utilisation et à la spécialisation de modèles de traduction automatique dans le cadre de projets de traduction, de post-édition et de révision. Équipée de cinq GPU, PNS-UP permet de réaliser des expériences dans le cadre de plusieurs projets de recherche et d'enseignement [20, 40, 41, 42]. Nous avons ainsi mis à disposition des petits et moyens modèles via le système Open Web UI, des accès GPU via des carnets Jupyter ou SSH.

Analyse des modèles de la parole

Au sein de l'équipe APREM, une attention particulière est accordée aux biais de tokenisation [24] dans les modèles multilingues, sur l'impact de l'algorithme BPE (*Byte Pair Encoding*) sur les dictionnaires de tokens des modèles , ainsi que sur les décalages entre les entrées des modèles de parole et leurs sorties, en exploitant notamment les représentations en sortie de l'encodeur des modèles Whisper.

Plus généralement, les travaux en explicabilité des grands modèles de langue portent sur les problématiques de la *learnability*, comment une annotation s'apprend [35], ce qu'un modèle de traduction neuronale apprend à chaque fois qu'il voit les données linguistiques [45], comment les grands modèles de langage peuvent contribuer à l'annotation.

Analyse des biais des modèles de langage

L'équipe APREM a réalisé des recherches sur l'explicabilité des modèles de TA neuronale [3] et de l'IA générative[41], notamment dans les cas de *fine-tuning* [31, 42]. L'équipe contribue à une réflexion globale sur les transformations qu'entraîne l'arrivée des modèles de langage à grande échelle pour les linguistes, notamment en ce qui concerne les procédures d'annotation [30]. Plus généralement, l'équipe conduit une réflexion épistémologique sur les nouvelles possibilités offertes par ces nouvelles technologies [2]. A cet effet, nous avons théorisé une analyse des sondes fondées sur les sorties des modèles, en particulier en traduction automatique neuronale [38] et en reconnaissance automatique de la parole, dans une forme d'ingénierie inversée [6] à une analyse plus classique des représentations fondée sur les matrices d'attention. Nous avons plus particulièrement travaillé avec des outils de visualisation, développés dans le cadre d'un projet financé par la Région Île-de-France (NeuroViz) et dans le cadre d'un autre projet, financé par les universités de Swansea et de Grenoble, MAKE-NMT-VIZ [21]. Pour la reconnaissance automatique de la parole, nous avons mis en évidence des problématiques de calibrage de modèles pour Whisper [6], ainsi que des biais de tokenisation. Nous avons approfondi cette problématique pour les langues moins bien dotées dans le cadre d'une collaboration financée par l'Université de Washington GIFRA [24], et dans le cadre d'un projet collaboratif soumis à la *National Science Foundation*. En collaboration avec le LLF, des travaux sont conduits reposant sur l'analyse des sondes (complétés par une analyse fondée sur les sorties des modèles). Nous avons mis en évidence les biais de genre des modèles de traduction [39].

Autres projets : exemples

Dans le projet ENSUPECO, l'équipe a étudié l'intégration des plateformes de traduction automatique neuronale (TAN) dans l'enseignement, mettant en lumière les défis persistants liés à la gestion des phénomènes discursifs complexes et à l'interaction homme-machine, tout en proposant des pistes pour une hybridation pédagogique [20, 18].

ALTAE est un des partenaires de l'ANR Matos. Le projet MaTOS (*Machine Translation for Open Science*) vise à développer de nouvelles méthodes pour la traduction automatique (TA) intégrale de documents scientifiques, ainsi que des métriques automatiques pour évaluer la qualité des traductions produites.

Citons également le Partenariat Hubert Curien franco-tunisien, lauréat de l'appel à projets PHC Utique 2026 : Discours du Développement Durable : Traitement Informatique Phraséologique (D3TIP). Sur le plan applicatif, le projet D3TIP prévoit le développement d'un outil d'aide à la décision assisté par l'IA, intégrant un module interactif de requêtage fondé sur des données expertes validées, une base terminologique multilingue du développement durable et un outil d'exploitation de corpus dédié au projet.

Science Ouverte

Via le consortium CORLI (Corpus et Interactions), labellisé par l'infrastructure de recherche nationale Huma-Num, ALTAE participe activement, notamment à travers des formations, à la production et au traitement de corpus numériques écrits, oraux ou multimodaux. Nous avons notamment proposé une formation consacrée à l'usage de petits modèles pour l'annotation de données linguistiques. Les membres d'ALTAE contribuent aux groupes de travail du consortium, tel que le groupe Multilinguisme, qui se concentre sur les corpus multilingues et plurilingues et l'usage de l'IA (Axe de travail 1 – IA, LLMs et leurs usages (annotation, multilinguisme...). Par le dépôt régulier de nos productions scientifiques sur Nakala, ou d'autres espaces de dépôts de recherche, nous travaillons dans une perspective de Science Ouverte permettant la mutualisation des ressources et la reproductibilité des résultats [4].

Analyse automatique des productions d'apprenants

Les membres de l'équipe (Phonétique, Acquisition & Psycholinguistique) ont développé un ensemble de compétences autour de l'analyse des traces numériques clavier (*keylogging*). Ces compétences recouvrent un savoir-faire méthodologique en termes d'analyse des données en temps réel [11], de jets textuels et de la constitution de métriques d'analyse pour la prédiction du niveau des apprenants [1], notamment dans une perspective d'Environnements Informatiques pour l'Apprentissage Humain (EIAH) [8]. Dans le cadre d'un projet financé, en collaboration avec Londres (KCL) et Cambridge (ALTA), un corpus d'essais d'apprenants et de natifs, (et leur *keylogging*), annotés en niveaux du CEFR (Cadre européen commun de référence pour le français), a été déposé sur HuggingFace [37].

Par ailleurs, l'une des compétences historiques d'ALTAE est la constitution de corpus d'apprenants (corpus Longitudinal Diderot Longdale, corpus MELLANGE et sa typologie d'erreurs) et son analyse automatique [15], marquée notamment par la co-organisation d'une tâche de prédiction du niveau des apprenants en fonction de la complexité des métriques textuelles (compétition « my taylor is rich » [7] de la Conférence sur l'Apprentissage automatique (CAp 2018), un partenariat Hubert Curien avec l'Irlande [9], la participation à l'ANR Analytics for Language Learning [25] et la contribution à l'analyse automatique des écrits des apprenants dans les EIAH (environnements informatiques pour l'apprentissage humain) sous forme de plate-forme web [34], puis de micro-services intégrés à Moodle. L'analyse automatique de la complexité des textes a également été appliquée à l'analyse du discours de spécialité dans le cadre d'une thèse au laboratoire [14] et des travaux connexes sur l'exploration des modèles d'intelligence artificielle générative comme assistants de reformulation de contenu dans des contextes multilingues [47].

Ces dernières années, nous avons également développé des solutions d'analyse automatique de *scoring* de la prononciation de l'anglais [10], dans le cadre d'une méthodologie applicable à d'autres langues [5]. L'équipe a, entre autres, développé une expertise autour des modèles de Whisper, également utilisés dans le cadre de la détection des problèmes cognitifs [33]. Le *scoring* automatique est également utilisé dans le cadre de l'ANR SARA (Sound Articulations Reinforce Acquisition) pour mesurer l'amélioration de la prononciation (à partir de l'audio) suite à un entraînement articulatoire en ultrason avec retour visuel. Enfin, faisant fonds sur des travaux en apprentissage automatique [26], l'équipe a également utilisé la reconnaissance d'images pour l'analyse des images ultrasons pour l'analyse des réalisations phonétiques, notamment en situation d'apprentissage, mais en comparaison avec les réalisations des natifs [22]. L'IA sert notamment au débruitage des images IRM, à la segmentation anatomique ou au tracking des contours de la langue à partir d'image échographique. Dans le cadre du projet SEPALE (Solutions pour l'Enseignement de la Phonétique Appliquée aux Langues Etrangères), des enregistrements d'étudiants ont été réalisés pour des tâches de classification automatique d'accents [17]. Ces problématiques de classification automatique des locuteurs [16, 19] et des accents sont au coeur du projet PARAAF (Perception de l'Anglais et de Reconnaissance Automatique d'Accents à la Fac).

Membres impliqués

- Nicolas BALLIER (PR)
- Brice BRICAUD (IGE)
- Emmanuel FERRAGNE (PR)
- Nicolas FROELIGER (PR)
- Natalie KÜBLER (PR)
- Christopher GLEDHILL (PR)
- Hannah KING (MCF)
- Manon LELANDAIS (MCF)
- Cameron MORIN (MCF)
- Stéphane PATIN (PR)
- Mojca PECMAN (PR)
- Gabriel THIBERGE (IGR)
- Cristian VALDEZ (MCF)
- Jean-Baptiste YUNÈS (MCF HDR, IRIF (UMR 8243), collaboration)
- Lichao ZHU (MCF)
- Maria ZIMINA-POIROT (MCF)

Références

[1] Ahood Al Sawar, Erin Pacquetet, Cyriel Mallart, Andrew Simpkin, and Nicolas Ballier. Analyse exploratoire des traces numériques clavier pour la prédiction des niveaux d'apprenants. In Frédéric Bechet, Adrian-Gabriel Chifu, Karen Pinel-sauvagnat, Benoit Favre, Eliot Maes, and Diana Nurbakova, editors, *Actes de l'atelier Traitement de données langagières dynamiques par les outils et méthodes du TAL 2025 (DYN-TAL)*, pages 7–11, Marseille, France, 6 2025. ATALA.

[2] Taylor Arnold, Nicolas Ballier, Paula Lissón, and Lauren Tilton. Beyond lexical frequencies : using R for text analysis in the digital humanities. *Language Resources and Evaluation*, 53(4) :707–733, 2019.

[3] Nicolas Ballier, Nabil Amari, Laure Merat, and Jean-Baptiste Yunès. The learnability of the annotated input in NMT replicating (vanmassenhove and way, 2018) with OpenNMT. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5631–5640, Marseille, France, May 2020. European Language Resources Association.

[4] Nicolas Ballier, Nabil Amari, Laure Merat, and Jean-Baptiste Yunès. The learnability of the annotated input in NMT replicating (Vanmassenhove and Way, 2018) with OpenNMT. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5631–5640, 2020.

[5] Nicolas Ballier, Taylor Arnold, Adrien Méli, Tori Thurston, and Jean-Baptiste Yunès. Whisper for L2 speech scoring. *International Journal of Speech Technology*, 27(4) :923–934, 2024.

[6] Nicolas Ballier, Léa Burin, Behnoosh Namdarzadeh, Sara B Ng, Richard Wright, and Jean-Baptiste Yunès. Probing whisper predictions for French, English and Persian transcriptions. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 129–138, 2024.

[7] Nicolas Ballier, Stéphane Canu, Caroline Petitjean, Gilles Gasso, Carlos Balhana, Theodora Alexopoulou, and Thomas Gaillat. Machine learning for learner English : A plea for creating learner data challenges. *International Journal of Learner Corpus Research*, 6(1) :72–103, 2020.

[8] Nicolas Ballier, Thomas Gaillat, and Erin Pacquetet. Prototype de feedback visuel des productions écrites d'apprenants francophones de l'anglais sous moodle. In *Environnements Informatiques pour l'Apprentissage Humain (EIAH'19)*, 2019.

[9] Nicolas Ballier, Thomas Gaillat, Andrew Simpkin, Bernardo Stearns, Manon Bouyé, and Manel Zarrouk. A supervised learning model for the automatic assessment of language levels based on learner errors. In *European Conference on Technology Enhanced Learning*, pages 308–320. Springer, 2019.

[10] Nicolas Ballier and Adrien Méli. Investigating Acoustic Correlates of Whisper Scoring for L2

Speech Using Forced Alignment with the Italian Component of the ISLE corpus. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 20–32, 2024.

[11] Nicolas Ballier, Erin Pacquetet, and Taylor Arnold. Investigating keylogs as time-stamped graphemics. In *Graphemics in the 21st Century 2018*, pages 353–365, 2018.

[12] Nicolas Ballier and Maria Zimina-Poirot. Littératie de la traduction automatique (TA) neuronale et traduction spécialisée : s' approprier les outils de la TA au travers de projets de recherche interdisciplinaires. *Human Translation and Natural Language Processing Towards a New Consensus ?*, 2023. Edizioni Ca'Foscari.

[13] Maud Bénard, Natalie Kübler, Alexandra Mestivier, Joachim Minder, and Lichao Zhu. Étude des Protocoles d' Évaluation Humaine pour la Traduction de Documents. Technical report, Projet ANR MaTOS, September 2024.

[14] Manon Bouyé. *Le style clair en droit : étude comparative du discours juridique en anglais et en français, avant et après simplification. Le plain language dans la communication juridique avec le grand public.* PhD thesis, Université Paris Cité, 2022.

[15] Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, editors. *Automatic Treatment and Analysis of Learner Corpus Data*, volume 59 of *Studies in Corpus Linguistics*. John Benjamins, Amsterdam, 2013.

[16] Emmanuel Ferragne, Cédric Gendrot, and Thomas Pellegrini. Towards phonetic interpretability in deep learning applied to voice comparison. In *ICPhS Proceedings*, pages 790–794, 2019.

[17] Emmanuel Ferragne, Anne Guyot Talbot, Hannah King, and Sylvain Navarro. Exploring the accent mix perceptually and automatically : French learners of English and the RP–GA divide. *Languages*, 9(2) :50, 2024.

[18] Nicolas Froeliger, Christopher Gledhill, and Maria Zimina-Poirot. Intégrer des plateformes de traduction automatique neuronale dans l'enseignement de la traduction spécialisée. *Traduction humaine et traitement automatique des langues-Vers un nouveau consensus ?*, 2023. Edizioni Ca Foscari.

[19] Cédric Gendrot, Emmanuel Ferragne, and Anaïs Chanclu. Analyse phonétique de la variation inter-locuteurs au moyen de réseaux de neurones convolutifs : voyelles seules et séquences courtes de parole. In *Journées d'étude de la parole 2022 (JEP 2022)*, 2022.

[20] Christopher Gledhill and Maria Zimina-Poirot. The Impact of Machine Translation on a Masters Course in Web Translation : From Disrupted Practice to a Qualitative Translation/Revision Workflow. In *Translating and the Computer 41*, pages 60–73, London, United Kingdom, November 2019. AsLing, The International Association for Advancement in Language Technology.

[21] Gabriela Gonzalez-Saez, Fabien Lopez, Mariam Nakhle, James Turner, Nicolas Ballier, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Caroline Rossi, Didier Schwab, and Jun Yang. The MAKE-NMTViz project : Meaningful, accurate and knowledge-limited explanations of NMT systems for translators. In Carolina Scarton, Charlotte Prescott, Chris Bayliss, Chris Oakley, and Joanna Wright, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 12–13, Sheffield, UK, June 2024. European Association for Machine Translation (EAMT).

[22] Hannah King and Emmanuel Ferragne. Labiodentals/r/here to stay : Deep learning shows us why. *Anglophonia. French Journal of English Linguistics*, (30), 2020.

[23] Claire Larsonneur, Nicolas Froeliger, and Giuseppe Sofo. *Human Translation and Natural Language Processing : Towards a new consensus ?* Edizioni Ca Foscari, December 2023.

[24] Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. The limits of data scaling : Sub-token utilization and acoustic saturation in multilingual ASR. In *Language Resources and Evaluation Conference*, 2026, to appear.

[25] Cyriel Mallart, Andrew Simpkin, Nicolas Ballier, Paula Lissón, Rémi Venant, Bernardo Stearns, Jen-Yu Li, and Thomas Gaillat. Assessing the validity of syntactic alternations as criterial features of proficiency in L2 writings in English. *Research Methods in Applied Linguistics*, 4(3) :100238, 2025.

[26] Adrien Méli and Nicolas Ballier. Analyse de la production de voyelles anglaises par des apprenants francophones, l'acquisition du contraste /l/-/i/ a la lumière des k-nn. *Anglophonia. French Journal of English Linguistics*, (27), 2019.

[27] Joachim Minder, Guillaume Wisniewski, and Natalie Kübler. Testing LLMs' capabilities in annotating translations based on an error typology designed for LSP translation : First experiments with ChatGPT. In Pierrette Bouillon, Johanna Gerlach, and Sabrina Girletti, editors, *Proceedings of Machine Translation Summit XX : Volume 1*, pages 190–203, Geneva, Switzerland, June 2025. European Association for Machine Translation.

[28] Joachim Minder, Guillaume Wisniewski, and Natalie Kübler. Testing LLMs' Capabilities in Annotating Translations Based on an Error Typology Designed for LSP Translation : First Experiments with ChatGPT. In *Proceedings of Machine Translation Summit XX : Volume 1*, pages 190–203, 2025.

[29] Cameron Morin and Matti Marttinen Larsson. A large-scale pipeline for automatic corpus annotation using LLMs : variation and change in the English consider construction, 2026.

[30] Cameron Morin and Matti Marttinen Larsson. Large corpora and large language models : a replicable method for automating grammatical annotation. *Linguistics Vanguard*, 11(1) :501–510, 2025.

[31] Behnoosh Namdarzadeh, Sadaf Mohseni, Lichao Zhu, Guillaume Wisniewski, and Nicolas Ballier. Fine-tuning MBART-50 with French and Farsi data to improve the translation of Farsi dislocations into English and French. In Masaru Yamada and Felix do Carmo, editors, *Proceedings of Machine Translation Summit XIX, Vol. 2 : Users Track*, pages 152–161, Macau SAR, China, September 2023. Asia-Pacific Association for Machine Translation.

[32] Mojca Pecman. Drafting definitions for emerging concepts and terms undergoing semantic shiftwithin the ARTES knowledge base : A protocol for integrating LLMs into terminological analysis by experimental approach. *Terminologija*, 32 :1–43, 2025.

[33] Amélie B. Richard, Manon Lelandais, and Sophie Jacquin-Courtois. Assessing cancer-related cognitive impairment for breast cancer survivors with speech analysis. *Speech Communication*, 174 :103284, 2025.

[34] Annanda Sousa, Nicolas Ballier, Thomas Gaillat, Bernardo Stearns, Manel Zarrouk, Andrew Simpkin, and Manon Bouyé. From linguistic research projects to language technology platforms : A case study in learner data. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 112–120, 2020.

[35] Zineddine Tighidet and Nicolas Ballier. Fine-tuning a subtle parsing distinction using a probabilistic decision tree : the case of postnominal "that" in noun complement clauses vs. relative clauses. In Pradeesh Parameswaran, Jennifer Biggs, and David Powers, editors, *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*, pages 52–61, Adelaide, Australia, December 2022. Australasian Language Technology Association.

[36] Cristian Valdez and Maria Lomeña Galiano. *Traducción automática neuronal en el ámbito de las lenguas romances. Potencial, límites y retos*. Tirant lo Blanch, October 2025.

[37] Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. Logging keystrokes in writing by English learners. In Nicoletta Calzolari, Min-Yen Kan, and Veronique Hoste, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics,*

Language Resources and Evaluation (LREC-COLING 2024), pages 10725–10746, Torino, Italia, May 2024. ELRA and ICCL.

[38] Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. Analyzing gender translation errors to identify information flows between the encoder and decoder of a nmt system. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 153–163, 2022.

[39] Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. Biais de genre dans un système de traduction automatique neuronale : une étude des mécanismes de transfert cross-langue. In *Traitement Automatique des Langues, Volume 63, Numéro 1 : Varia [Varia]*, pages 37–61, 2022.

[40] Lichao Zhu, Maria Zimina, Maud Bénard, Behnoosh Namdar, Nicolas Ballier, Guillaume Wisniewski, and Jean-Baptiste Yunès. Investigating techniques for a deeper understanding of neural machine translation (NMT) systems through data filtering and fine-tuning strategies. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 275–281, Singapore, December 2023. Association for Computational Linguistics.

[41] Lichao Zhu, Maria Zimina, Behnoosh Namdarzadeh, Nicolas Ballier, and Jean-Baptiste Yunès. Enhancing translation quality : A comparative study of fine-tuning and prompt engineering in dialog-oriented machine translation systems. insights from the MULTITAN-GML team. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1016–1022, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[42] Lichao Zhu, Maria Zimina-Poirot, Cristian Valdez, and Stephane Patin. Fine-tuning NMT models and LLMs for specialised EN-ES translation using aligned corpora, glossaries, and synthetic data : MULTITAN at WMT25 terminology shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 1284–1291, Suzhou, China, November 2025. Association for Computational Linguistics.

[43] Maria Zimina and Christopher Gledhill. Human-Machine Interaction : How to Integrate Plain Language Rules in the Revision Cycles of Neural Machine Translation (NMT) output. Conférence Linguistic Rights and Language Varieties in Europe in the Age of AI, Università di Torino. 2021.

[44] Maria Zimina-Poirot. Computer-assisted language mediation in teaching human-centred augmented translation. In *Translation and Multilingual Natural Language Processing 25*. Language Science Press, 2025.

[45] Maria Zimina-Poirot, Nicolas Ballier, and Jean-Baptiste Yunès. Approches quantitatives de l'analyse des prédictions en traduction automatique neuronale (TAN). In *JADT 2020 : 15èmes Journées Internationales d'Analyse statistique des Données Textuelles*, Toulouse, France, June 2020. Université de Toulouse.

[46] Maria Zimina-Poirot and Christopher Gledhill. Comment les étudiants révisent-ils les sorties de la traduction automatique neuronale ? Quelques réflexions sur l'application des directives de la DGT et l'impact de l'anglais lingua franca sur les processus de traduction.# Translating Europe Workshop L'intelligenza artificiale per la traduzione : verso una nuova progettazione didattica ? 2021.

[47] Maria Zimina-Poirot, Christopher Gledhill, and Manon Bouyé. Towards Clear, Simple, and Inclusive Legal Language : Exploring Generative Artificial Intelligence (GAI) Models as Content Rephrasing Assistants in Multilingual Settings. *International Journal of Language & Law (JLL)*, 14 :143–173, 2025.

[48] Maria Zimina-Poirot, Christopher Gledhill, and Manon Bouyé. Towards clear, simple, and inclusive legal language — exploring generative artificial intelligence (GAI) models as content rephrasing assistants in multilingual settings. *International Journal of Language & Law (JLL)*, 14 :143–173, May 2025.

AfIA
Association française
pour l'Intelligence Artificielle

AMIAD/Pôle Recherche : Groupe langage

Pôle Recherche : Groupe langage
AMIAD
https://www.amiad.fr

Jean-François BONASTRE
Responsable du groupe
jean-francois.bonastre@polytechnique.edu

Thématiques

Créé le 1/10/2026, le groupe "Langage et audio" du Pôle Recherche de l'Agence Ministérielle pour l'IA de Défense (AMIAD) s'intéresse principalement aux technologies de la langue dans le cadre des applications Défense. Le groupe mène également des travaux en Recherche d'Information (RI) et sur l'audio. Les principales thématiques abordées sont :

- Reconnaissance Automatique de la Parole
- Reconnaissance/Identification du locuteur et de la langue
- Traduction de la Parole
- Communication Humain–Machine
- Description automatique de scènes audio
- Méthodes d'évaluation pour le TAL et la RI
- Utilisabilité des Grands Modèles de Langues (LLM)
- Langues peu dotées
- Robustesse des traitements et adaptation de domaine

Travaux en cours

Traduction de la parole

Le projet vise à apprendre ou à adapter un système de traduction de la parole vers du texte (en français) de manière frugale, tant en données brutes qu'en données transcrites. Le but recherché est de rendre un système de traduction opérationnel pour une nouvelle langue très rapidement, même si cette langue est très peu dotée en ressources.

Reconnaissance de la langue maternelle dans une production langagière

Les travaux visent à développer un système capable de reconnaître la langue maternelle d'une personne à partir d'une production de celle-ci réalisée dans une autre langue. L'approche retenue s'appuie sur des LLM. Les premiers travaux portent sur des productions écrites, et une extension du texte à la parole est envisagée.

Détection de deepfakes audio

Les travaux visent à développer un système de détection des deepfakes audio. Contrairement aux solutions actuelles, l'approche retenue consiste à rechercher les anomalies par rapport à un enregistrement "naturel". Deux objectifs sont recherchés à travers cette approche : être capable de généraliser à des générateurs de deepfake jamais vus durant l'apprentissage et permettre à l'opérateur d'ajuster la sensibilité de détection en fonction du contexte, de façon analytique et transparente, type d'anomalie par type d'anomalie.

Attaque de détecteurs de deepfakes audio

Les détecteurs de deepfakes audio (DDA) sont sensibles à la qualité de l'outil de génération utilisé mais aussi à des attaques adversariales qui adaptent les audios générés pour mieux tromper un DDA spécifique. Nous concevons une telle attaque adversariale sous la forme d'une perturbation de type filtre de convolution, qui offre des avantages en termes de généricité, de simplicité et d'explicabilité.

Détection et clustering de la langue parlée

Les travaux visent à développer un système de reconnaissance de la langue parlée explicable reposant sur une approche par attributs, ou concepts, qui décompose la décision en contributions d'attributs permettant de caractériser les langues. Un tel système vise également à faciliter l'ajout de nouvelles langues non présentes dans le corpus d'entraînement et à caractériser des langues inconnues.

Apprentissage frugal et adaptation de domaine

Ces travaux étudient l'adaptation de domaine dans un régime où une faible quantité de données du domaine cible est disponible pour l'adaptation, associée à des ressources de calcul potentiellement très limitées. L'application visée consiste à adapter un système de réhaussement de la parole à de nouvelles conditions acoustiques.

Vérification du locuteur cross-domaine

La vérification du locuteur consiste à comparer deux enregistrements audio afin de déterminer s'ils ont été prononcés par le même locuteur. Ce travail vise à modéliser et neutraliser au maximum l'incertitude due à une variation de domaine entre les deux enregistrements. Les variations de domaine explorées sont le canal de transmission et la langue.

LLM fondation et frugalité orienté données

L'objectif premier des travaux est de maîtriser la chaîne de création des LLM génératifs de A à Z. Le deuxième objectif concerne la frugalité, avec des travaux visant à être beaucoup moins gourmands en données pour l'apprentissage des LLM.

Traçabilité des productions des LLM

L'objectif est la traçabilité et la validation des générations d'un LLM, par la vérification de la cohérence entre le texte généré et les sorties issues de procédés de recherche d'information (comme la RAG). Une approche de type simplification de texte par la décomposition des réponses générées en « propositions atomiques » (phrases minimales) a été retenue, ce qui permet d'identifier plus aisément les hallucinations.

Évaluation des LLM

Afin de mesurer l'efficacité des approches LLM, l'évaluation est devenue une priorité, notamment pour contrer l'effet de contamination : la présence potentielle des données d'évaluation dans les données d'apprentissage ou de pré-apprentissage des LLM. Nos travaux sur le sujet explorent les méthodes et les moyens de détecter les contaminations, de les extraire et de s'en protéger.

Membres permanents

- Olivier Boeffard
- Raphaël Duroselle
- Christophe Servan
- Thomas Thebaud

Références

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0 : A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[2] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret 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*, pages 610–623, 2021.

[3] Anupama Chadha, Vaibhav Kumar, Sonu Kashyap, and Mayank Gupta. Deepfake : an overview. In *Proceedings of second international conference on computing, communications, and cyber-security : IC4S 2020*, pages 557–566. Springer, 2021.

[4] Louis Estève, Christophe Servan, Thomas Lavergne1, and Agata Savary1. A diversity diet for a healthier model : A case study of french modernbert. In *Proceedings of the Findings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, 2026.

[5] Dan Feldman. Core-sets : Updated survey. In *Sampling techniques for supervised or unsupervised tasks*, pages 23–44. Springer, 2019.

[6] Dhiman Goswami, Jai Kruthunz Naveen Kumar, and Sanchari Das. Nlp privacy risk identification in social media (nlp-prism) : A survey. In *Findings of the Association for Computa-*

tional Linguistics : EACL 2026, pages 1519–1541, 2026.

[7] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli-Barone, Jindřich Helcl, and Alexandra Birch. Survey of low-resource machine translation. *Computational Linguistics*, 48(3) :673–732, 2022.

[8] Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. Deepfakes : Trick or treat ? *Business horizons*, 63(2) :135–146, 2020.

[9] Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, and Laurent Besacier. Speech-MASSIVE : A Multilingual Speech Dataset for SLU and Beyond. In *Proc. Interspeech 2024*, 2024.

[10] Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, and Sophie Rosset. Détection des contaminations de llm par extraction de données : une revue de littérature pratique. In *Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, pages 233–251, 2025.

[11] Shervin Malmasi, Joel Tetreault, and Mark Dras. Oracle and human baselines for native language identification. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 172–178, 2015.

[12] Yee Man Ng and Ilia Markov. Leveraging open-source large language models for native language identification. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28, 2025.

[13] Terence Odlin. Cross-linguistic influence. *The handbook of second language acquisition*, pages 436–486, 2003.

[14] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages : A survey. *ACM Computing Surveys*, 55(11) :1–37, 2023.

[15] Monica T Whitty. Drug mule for love. *Journal of Financial Crime*, 30(3) :795–812, 2023.

[16] Wei Zhang and Alexandre Salle. Native language identification with large language models. *arXiv preprint arXiv :2312.07819*, 2023.

AfIA
Association française
pour l'Intelligence Artificielle

MaIAGE/BIBLIOME : Acquisition et Formalisation de Connaissances à partir de Textes

MaIAGE UR 1404
INRAE et Université Paris-Saclay
https://maiage.inrae.fr/bibliome

Robert BOSSY
robert.bossy@inrae.fr

Claire NÉDELLEC
claire.nedellec@inrae.fr

Louise DELÉGER
louise.deleger@inrae.fr

Arnaud FERRÉ
arnaud.ferre@inrae.fr

Domaine de recherche

L'équipe Bibliome développe des méthodes d'extraction et de formalisation d'information à partir de textes écrits. Ces méthodes identifient et structurent des informations et connaissances dans de larges corpus de documents hétérogènes, en mobilisant des approches de traitement automatique de la langue, de l'apprentissage neuronal profond et des modèles de langue.

Les travaux de l'équipe s'organisent autour de l'exploitation de modèles de langue pré-entraînés (modèles BERT ou ses variantes) et de grands modèles de langue (LLM), ainsi que de l'intégration de représentations construites à partir de bases de connaissances. Les principaux travaux concernent :

1. l'apprentissage neuronal profond pour la reconnaissance, la normalisation et la mise en relation d'entités ;
2. la conception, l'alignement et l'exploitation de terminologies, ontologies et graphes de connaissances ;
3. l'évaluation, la robustesse et l'adaptation des modèles dans des contextes de faible supervision et de transfert de domaine (*few-shot/zero-shot*).

Ces travaux de recherche s'inscrivent principalement dans une dynamique d'exploitation de la littérature scientifique, dans un objectif d'analyse des connaissances qui y sont décrites. Les sciences du vivant, l'agriculture, l'alimentation et la santé constituent des terrains d'étude privilégiés pour le développement, l'analyse et la validation de nouvelles méthodes en extraction et structuration de connaissances.

Contributions méthodologiques

Les méthodes développées par l'équipe Bibliome couvrent l'extraction d'entités nommées, leur normalisation par des concepts d'ontologies, l'extraction de relations entre ces entités, et la fiabilité de ces informations. Ces travaux s'inscrivent dans des contextes où les données annotées sont rares, coûteuses à produire et fortement dépendantes de l'expertise métier, et visent à mieux comprendre les limites des modèles actuels.

Depuis plus de 10 ans, les travaux de l'équipe Bibliome ont fortement intégré les approches neuronales et les modèles de langue pré-entraînés. Ils ont notamment étudié l'apport de modèles de type BERT enrichis par des informations syntaxiques [17, 18]. Ils explorent également l'intégration de plongements de graphes pour améliorer l'extraction de relations [16].

L'équipe développe des méthodes robustes pour la reconnaissance d'entités dans des contextes bruités, comme l'approche SCoNE basée sur l'autocorrection et l'augmentation de données [22]. Des travaux portent également sur l'adaptation de domaine et l'amélioration des modèles par des stratégies d'apprentissage à base de bi-encoders [1].

La normalisation d'entités (ou liage d'entités - *entity linking*) constitue un axe important de leurs travaux, en raison de son rôle clé dans l'articulation entre extraction textuelle et représentation des

connaissances. Ils développent notamment des approches neuronales dédiées combinant représentations vectorielles et structures ontologiques, telles que C-Norm [3], et mènent des analyses approfondies des biais d'évaluation des méthodes de l'état de l'art [4].

Dans la continuité de ses travaux sur les représentations distributionnelles puis sur les modèles de langue contextualisés, l'équipe explore aussi depuis plusieurs années l'usage des LLMs, notamment pour l'extraction d'information et l'évaluation de la cohérence des prédictions [20], leur intégration avec des approches de type RAG pour la vérification factuelle [21] et l'analyse de la fiabilité des informations à travers l'étude des fonctions rhétoriques des citations et des discours académiques représentés sous forme de graphe [5].

L'équipe amorce également des travaux sur l'intégration de données multimodales (textes, mais aussi images, tableaux, graphes ou encore séquences biologiques), afin d'étudier l'apport de ces différentes sources pour l'extraction et la structuration des connaissances.

Domaine d'application

Les domaines d'application de l'équipe couvrent un large spectre des sciences du vivant, incluant la microbiologie, la biologie végétale, la santé des plantes, l'écologie et l'agriculture. Les travaux menés dans ces domaines constituent également des cas d'étude pour l'analyse et la validation des méthodes développées.

Les graphes de connaissances constituent un cadre privilégié pour l'intégration et la structuration d'informations issues de la littérature scientifique, notamment en génomique du blé et du riz [19]. L'équipe contribue également à des problématiques de santé globale (*One Health*) appliquées à la santé des plantes [7] dans un contexte de changement climatique [15].

Les applications incluent l'extraction d'informations sur la biodiversité microbienne (ex. Omnicrobe), les interactions biotiques et abiotiques (ex. [10]), les phénotypes et traits agronomiques (ex. WheatIS Data Discovery), ainsi que l'analyse de la littérature scientifique.

Construction de ressources sémantiques partagées

L'équipe conçoit, produit et diffuse des ressources sous licence ouverte, notamment des corpus annotés et des ontologies. Elle organise régulièrement des compétitions internationales, dont la tâche PestCLEF du Lab LifeCLEF [6] basée sur le nouveau corpus EPOP (Epidemiomonitoring of Plant) pour l'extraction de graphes d'information en santé des plantes [9].

Ces dernières années, elle a développé plusieurs nouvelles ressources de référence, telles que le corpus TaeC pour l'extraction de traits et phénotypes dans la littérature sur le blé [12], des corpus annotés pour des domaines spécifiques comme HoloOligoCorpus [13], ainsi que des ressources lexicales spécialisées comme MilkOligoThesaurus [14] ou Wheat Trait Ontology.

L'équipe travaille également sur l'interopérabilité sémantique via l'alignement d'ontologies, par exemple pour les traits du blé [11, 8]. Ces travaux s'inscrivent dans une démarche de "données ouvertes liées" et de science ouverte.

Les ressources produites sont utilisées pour entraîner et évaluer des modèles d'apprentissage profond et pour construire des graphes de connaissances exploitables à grande échelle. Elles sont distribuées et documentées sous licence ouverte, et peuvent être valorisées à travers des services finalisés permettant leur exploration à grande échelle, tels que la plateforme Omnicrobe [2]

Projets

L'équipe Bibliome est impliquée dans plusieurs projets collaboratifs nationaux et européens, portant sur l'extraction et la structuration de connaissances à partir de données textuelles dans les sciences du vivant, ainsi que sur l'intégration de méthodes récentes en apprentissage neuronal profond et en traitement automatique des langues.

Parmi ces projets, le réseau doctoral européen FAIROmics (2023-2027) vise à structurer et relier des données multi-omiques pour la construction de graphes de connaissances dans le domaine des aliments fermentés. L'équipe s'implique dans la transition agroécologique à travers le projet BEYOND (2020-2026), financé dans le cadre du programme «

Cultiver et protéger autrement », sur la surveillance et l'anticipation du risque épidémiologique en santé des plantes à partir de données hétérogènes. Dans les projets ANR EcoControl (2025-2029) et partenarial « Facteurs de Succès » (2024-2027) sur le développement de méthodes de biocontrôle, sa contribution porte sur l'enrichissement incrémental de graphes de connaissance à partir d'articles scientifiques. Récemment, plusieurs projets ont débuté sur l'exploration de l'usage des LLMs et des approches hybrides. Le projet AIKO porte sur l'extraction de connaissances à partir de publications scientifiques, tandis que le projet KPREL (2025-2026) étudie l'intégration de bases de connaissances dans des approches de type *prompting* pour l'extraction de relations.

L'équipe est également impliquée dans des projets applicatifs en santé et en agriculture, tels que l'ANR SafeFood4ClimDiet (2025-2027), qui vise à analyser les risques microbiologiques liés aux évolutions des pratiques alimentaires induites par le changement climatique.

Au niveau national, l'équipe contribue régulièrement à des projets structurants de formation autour de l'extraction d'information et de l'intelligence artificielle, notamment à travers l'Action Nationale de Formation (ANF) « TDM & IA » co-organisée par le CNRS et INRAE, ainsi que les Écoles Thématiques Interdisciplinaires ETLMIC, consacrées aux modèles de langue et à l'intégration de connaissances, qui participent à la diffusion des méthodes et à la structuration de la communauté scientifique à INRAE.

Références

[1] Mariya Borovikova, Arnaud Ferré, Robert Bossy, Mathieu Roche, and Claire Nédellec. Semantically-Informed Domain Adaptation for Named Entity Recognition. In Annalisa Appice, Hanane Azzag, Mohand-Said Hacid, Allel Hadjali, and Zbigniew Ras, editors, *LNAI-14670 Foundations of Intelligent Systems*, volume 14670 of *Lecture Notes in Computer Science*, pages 55–64, Poitiers, France, June 2024. Springer Nature Switzerland.

[2] Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba, Estelle Chaix, Olivier Harlé, Valentin Loux, Hélène Falentin, and Claire Nédellec. Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach. *PLoS ONE*, 18(1) :e0272473, January 2023.

[3] Arnaud Ferré, Louise Deléger, Robert Bossy, Pierre Zweigenbaum, and Claire Nédellec. C-norm : a neural approach to few-shot entity normalization. *BMC Bioinformatics*, 21(23) :1–19, 2020.

[4] Arnaud Ferré and Philippe Langlais. An analysis of entity normalization evaluation biases in specialized domains. *BMC Bioinformatics*, 24(1) :227, June 2023.

[5] Anne-Sophie Foussat, Vincent Guigue, Nicolas Sauvion, Robert Bossy, and Claire Nédellec. Transfert de modles de langue pour la classification rhétorique des citations
travers les disciplines. In Frédéric BECHET, Adrian-Gabriel CHIFU, Karen PINEL-SAUVAGNAT, Benoit FAVRE, Eliot MAES, and Diana NURBAKOVA, editors, *Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA)*, pages 233–248, Marseille, France, 2025. ATALA
ARIA. 32me Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 27me Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) Les 18e Rencontres Jeunes Chercheurs en RI (RJCRI).

[6] Alexis Joly, Lukáš Picek, Stefan Kahl, Hervé Goëau, Lukáš Adam, Robert Bossy, Kostas Papafitsoros, Vojtěch Čermák, Holger Klinck, Willem-Pier Vellinga, et al. Lifeclef 2026 teaser : Ai challenges for biodiversity understanding and ecosystem management. In *European Conference on Information Retrieval*, pages 287–296. Springer, 2026.

[7] Cindy E. Morris, Ghislain Geniaux, Claire Nédellec, Nicolas Sauvion, and Samuel Soubeyrand. One Health concepts and challenges for surveillance, forecasting and mitigation of plant disease beyond the traditional scope of crop production. *Plant Pathology*, 71 :86–97, 2022.

[8] Claire Nédellec, Sophie Aubin, Clara Sauvion, Liliana Ibanescu, Sonia Bravo, Jacques Le-Gouis, Thierry C Marcel, Cyril Pommier, Robert Bossy, and Michael Alaux. Mapping bread wheat trait ontologies for semantic interoperability. *F1000Research*, page 1102, September 2024.

[9] Claire Nédellec, Marine Courtin, Xinzhi Yao, Marie Grosdidier, Isabelle Pieretti, Sandy Dupérier, and Robert Bossy. EPOP : A benchmark corpus for Assessing NLP Models on Structured Information Extraction in Plant Health. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC)*, Palma, Mallorca, Spain, May 2026. May 11-16, 2026.

[10] Claire Nédellec, Marine Courtin, Xinzhi Yao, Marie Grosdidier, Isabelle Pieretti, Sandy Duperier, and Robert Bossy. Video of the presentation at lrec conference : Epop : A benchmark corpus for assessing nlp models on structured information extraction in plant health. 2026.

[11] Claire Nédellec, Liliana Ibanescu, Robert Bossy, and Pierre Sourdille. WTO, an ontology for wheat traits and phenotypes in scientific publications. *Genomics & Informatics*, 18(2), June 2020.

[12] Claire Nédellec, Clara Sauvion, Robert Bossy, Mariya Borovikova, and Louise Deléger. TaeC : A manually annotated text dataset for trait and phenotype extraction and entity linking in wheat breeding literature. *PLoS ONE*, 19(6) :e0305475, 2024.

[13] Mathilde Rumeau, Marine Courtin, Robert Bossy, Clara Sauvion, Valentin Loux, Mouhamadou Ba, Christelle Knudsen, Sylvie Combes, Claire Nédellec, and Louise Deléger. MilkOligoCorpus : A semantically annotated resource for knowledge extraction on mammalian milk oligosaccharides. *PLoS ONE*, 20(8) :e0319729, August 2025.

[14] Mathilde Rumeau, François Fenaille, Agns Girard, Valentin Loux, Mouhamadou Ba, Claire Nédellec, Louise Deleger, Robert Bossy, Sophie Aubin, Christelle Knudsen, and Sylvie Combes. MilkOligoThesaurus, a dataset of mammalian milk oligosaccharide synonyms. *Data in Brief*, 54 :110404, April 2024.

[15] S. Soubeyrand, Arnaud Estoup, Astrid Cruaud, Sylvie Malembic-Maher, Christine Meynard, Virginie Ravigné, Marc Barbier, Benoit Barrs, Karine Berthier, Simon Boitard, Sylvie Dallot, Sabrina Gaba, Marie Grosdidier, Mourad Hannachi, Marie-Agns Jacques, Melen Leclerc, Philippe Lucas, Davide Martinetti, Christophe Mougel, Corinne Robert, Alain Roques, Jean-Pierre Rossi, Frédéric Suffert, Pierre Abad, Marie-Anne Auger-Rozenberg, Jean-Sauveur Ay, Marc Bardin, Hélne Bernard, David Bohan, Thierry Candresse, Philippe Castagnone-Sereno, Etienne Danchin, Chloe E. L. Delmas, Pauline Ezanno, Frédéric Fabre, Benoît Facon, Edith Gabriel, Jonathan Gaudin, Bertrand Gauffre, Mathieu Gautier, Claire Guinat, Claire Lavigne, Olivier Lemaire, César Martinez, Lucie Michel, Benoît Moury, Kiwoong Nam, Claire Nédellec, Mylne Ogliastro, Julien Papaïx, Nicolas Parisey, Sylvain Poggi, Andrea Radici, Jean-Yves Rasplus, Xavier Reboud, Cecile Robin, Mathieu Roche, Adrien Rusch, Nicolas Sauvion, Jean-Claude Streito, Eric Verdin, Anne-Sophie Walker, Anne Xuéreb, Gaël Thébaud, and Cindy E. Morris. Building integrated plant health surveillance : a proactive research agenda for anticipating and mitigating disease and pest emergence. *CABI Agriculture and Bioscience*, 5 :72, August 2024.

[16] Anfu Tang, Robert Bossy, Louise Deléger, Claire Nédellec, and Pierre Zweigenbaum. Exploitation de plongements de graphes pour l'extraction de relations biomédicales. In Christophe Servan and Anne Vilnat, editors, *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*, volume 1 : travaux de recherche originaux – articles longs, pages 298–310, Paris, France, June 2023. ATALA.

[17] Anfu Tang, Louise Deleger, Robert Bossy, Pierre Zweigenbaum, and Claire Nédellec. Do syntactic trees enhance Bidirectional Encoder

Representations from Transformers (BERT) models for chemical–drug relation extraction ? *Database - The journal of Biological Databases and Curation*, 2022 :baac070, August 2022.

[18] Anfu Tang, Claire Nédellec, Pierre Zweigenbaum, Louise Deléger, and Robert Bossy. Global alignment for relation extraction in Microbiology. In *Proceedings of the Junior Conference on Data Science and Engineering 2021*, Proceedings of the Junior Conference on Data Science and Engineering 2021, Orsay, France, February 2021.

[19] Nadia Yacoubi Ayadi, Franck Michel, Robert Bossy, Marine Courtin, Bill Gates Happi Happi, Pierre Larmande, Claire Nedellec, and Catherine Faron. The Wheat and Rice Genomics Scientific Literature Knowledge Graphs. In *Semantic Web Technologies for Health Care and Life Science (SWAT4HCLS)*, Barcelone, Spain, February 2025. Andra Waagmeester and M. Scott Marshall and Andrea Splendiani.

[20] Xinzhi Yao, Claire Nédellec, Jingbo Xia, and Robert Bossy. Consistency–accuracy correlation in hard-prompted LLMs for entity and relation extraction : empirical findings from plant-health data. *Genomics & Informatics*, 24(1) :3, February 2026.

[21] Lidan Zhang, Louise Deléger, Olivier Inizan, and Arnaud Ferré. Evaluation of LLMs with RAG for Factual Verification in the Biomedical Domain. In *10th edition of the Junior Conference on Data Science and Engineering (JDSE)*, Gif-sur-Yvette, France, September 2025.

[22] Xingyu Zhu, Claire Nédellec, Balazs Nagy, Laszlo Vidacs, and Robert Bossy. SCoNE : a Self-Correcting and Noise-Augmented Method for Complex Biological and Chemical Named Entity Recognition. In *European Chapter of the Association for Computational Linguistics (EACL)*, Rabat (MOROCCO), Morocco, March 2026.

AfIA
Association française
pour l'Intelligence Artificielle

CEA-List/LASTI : Laboratoire Analyse Sémantique Texte Image

LASTI
CEA-List
https://kalisteo.cea.fr/index.php/ai/

Bianca VIERU
bianca.vieru@cea.fr

Présentation générale

Au sein de l'institut LIST du CEA, le Laboratoire d'Analyse Sémantique Texte et Image (**LASTI**) regroupe environ 30 collaborateurs (chercheurs, ingénieurs et doctorants) et développe des méthodes pour l'analyse et la production de contenus multimodaux associant texte et image.

Les travaux du laboratoire s'articulent autour de deux axes principaux : d'une part, l'extraction et la structuration d'information à partir de documents multimodaux non structurés ; d'autre part, la génération contrôlée intégrant des contraintes explicites définies par l'utilisateur. Un axe transverse vient les renforcer par des contributions en explicabilité des modèles, en raisonnement automatique et en adaptation des modèles.

Les travaux académiques sont largement diffusés en *open source* via Github ou HuggingFace. En aval, le laboratoire a vocation à transférer ses technologies à des partenaires industriels.

Principaux thèmes de recherche

Depuis le début des années 2000, le CEA-List travaille dans le domaine du traitement des documents multimodaux dans une perspective d'accès aux contenus et de structuration de connaissances en mettant l'accent sur le multilinguisme et la multimodalité (texte, image, parole). Ces travaux sont conduits en intégrant les contraintes propres aux contextes industriels, notamment la performance attendue et le maintien de cette performance dans le temps, l'adaptation à des données et tâches spécifiques, la disponibilité limitée de données spécifiques, les restrictions en ressources de calcul ainsi que la nécessité de maîtriser finement les processus de génération de contenu.

Structuration d'information multimodale

Extraction d'information. Le LASTI couvre un large spectre de tâches d'extraction d'information allant de la reconnaissance d'entités nommées à l'extraction d'événements en passant par l'extraction de relations et le liage référentiel, ceci dans une perspective multilingue, comme au sein des projets H2020 Starlight et HE ARIEN. Il considère ces tâches en intégrant des contraintes de frugalité liées au nombre d'exemples annotés et à la taille des modèles utilisés, problématique importante pour un projet tel que Data4Industry-X. Cette priorité a aussi été centrale dans l'évaluation EvalLLM où le LASTI a occupé deux années consécutives la *1ère place* en s'appuyant sur des approches contrastives (GLiNER) surpassant les grands LLM décodeurs [6, 5] pour l'extraction d'entités et d'événements. Ce même souci a sous-tendu les travaux du LASTI sur le modèle GLiDRE [4] d'extraction de relations au niveau du document, testé dans le cadre de l'évaluation TextMine'25 où le LASTI a par ailleurs obtenu une *2ème place* avec des modèles exploitant la quantité plus importante de données fournie aux participants. L'adaptation de modèles avec peu de données se retrouve aussi dans les travaux menés sur l'application du méta-apprentissage à l'extraction des arguments d'événements [29] ou sur la génération automatique d'annotation en événements pour des corpus de préentraînement [11], des travaux développés dans le PEPR SHARP.

Question-réponse multimodale fondée sur les connaissances. Le LASTI a renforcé sa focalisation sur la représentation multimédia des entités nommées en tant que coordinateur du projet ANR MEERQAT. Ce projet a produit ViQuAE [23], un jeu de données de référence comprenant 1,5 million d'entités et 3 700 questions, permettant de valider des modèles de fusion de média et l'apport de connaissances externes [2]. Aujourd'hui intégré

dans des benchmarks internationaux pour les modèles vision-langage (M3IT, MVBench), ViQuAE a servi de socle aux travaux du LASTI ayant amélioré l'état de l'art des systèmes de QR multimodaux, d'abord par le pré-entraînement auto-supervisé [1] de représentations multimodales d'entités, puis en définissant le premier modèle de RAG cross-modal pour cette tâche [3].

Génération contrôlée

Génération de contenu. Après avoir exploré des aspects fondamentaux du contrôle des modèles génératifs purement visuels tels que les GAN/VAE [28] et le désentremêlement des attributs sémantiques [12, 13], le LASTI investigue le contrôle par des prompts textuels complexes des modèles de diffusion ou de flots conditionnels [20, 21, 19]. Le contrôle de la génération de texte a été exploré dans le projet ANR CODEINE pour définir le contenu de comptes rendus médicaux synthétiques [22] en fonction de profils patient prenant la forme d'un ensemble d'attributs médico-démographiques [7].

Plus récemment, avec la proposition du jeu de données CaMiT [24] comportant 787K images de voitures sorties sur la période 2007–2023, le LASTI introduit la génération conditionnée par le temps, permettant notamment de travailler sur l'adaptation incrémentale.

Création de bases de données synthétiques et équitables. Le LASTI a proposé des approches pour générer des jeux de données d'entraînement synthétiques équilibrés démographiquement (genre, âge, origine ethnique) pour la vérification faciale. Une analyse statistique montre que cet équilibrage améliore l'équité sans nuire aux performances [18]. Via le PEPR SHARP, une méthode reproductible d'inférence démographique a été proposée pour permettre un audit fiable de l'équité [17].

Agents conversationnels. La génération de contenu est utilisée également dans le contexte des agents conversationnels [25]. L'assistant CoVA du LASTI, développé dans HE Cortex^2, détermine avec précision les intentions des participants à une réunion pour donner des informations ou services au moment adéquat grâce à son aspect agentique [9, 8]. Cet agent est aussi intégré à des systèmes robotiques pour offrir une interaction naturelle [27], le travail dans ce cas portant plus spécifiquement sur la modélisation de mémoires à court, moyen et long terme afin de pouvoir exploiter conjointement des connaissances symboliques (graphes de connaissances, graphes de scène 3D, etc.) et des modèles neuronaux, en particulier dans HE ASTIR.

IA de confiance

Détection et réduction des biais dans les LLM. Ces travaux visent à caractériser ce que les LLM encodent réellement — biais politiques, préférences culturelles, propriétés des hiérarchies lexicales — en utilisant les entités nommées comme sondes statistiques, indépendamment des modèles eux-mêmes. Le LASTI propose une nouvelle manière de mesurer les biais politiques dans les grands modèles de langue. Les résultats mettent en lumière des préférences idéologiques inattendues, variables selon la langue et la taille du modèle. Les biais sont significativement réduits en anonymisant les noms des politiciens [14]. Les chercheurs du LASTI ont mis en place le plus grand audit de biais de LLM à ce jour (1,9Md de points de données) avec une adaptation semi-automatique de l'approche à des domaines applicatifs [16]. Les approches proposées dans le cadre du challenge CLEF 2025 sur la détection de subjectivité multilingue ont permis d'obtenir la première place en arabe et en polonais [15].

Les LLM ne font pas que halluciner, ils omettent aussi des informations. Or, cela peut être grave, par exemple dans le domaine médical. Pour pouvoir explorer les omissions dans les résumés des textes médicaux générés par LLM, le LASTI propose des méthodes frugales de détection de celles-ci, fondées sur des plongements lexicaux statiques (FastText, Word2Vec) combinés à des modèles de détection d'anomalies, et validées par des chirurgiens sur des données françaises réelles [26].

Représentations interprétables. Dans le projet ANR STUDIES, les travaux du LASTI sur l'identification des concepts au sein des LLM, ainsi que la manière dont ils sont liés, ont permis d'avancer sur l'interprétabilité mécaniste de ces modèles [10] et ainsi de mieux comprendre le comportement des grands modèles multimodaux (texte, image, audio), avec des applications en génération de données.

Membres permanents

- Bianca VIERU (responsable)
- Romaric BESANÇON
- Gaël de CHALENDAR
- Anne-Laure DAQUO
- Jérôme DESHAYES-CHOSSART
- Olivier FERRET (HDR)
- Sandra KARA
- Benjamin LABBÉ
- Meriama LAIB
- Hervé LE BORGNE (HDR)
- Adrian POPESCU (HDR)
- Nasredine SEMMAR (HDR)
- Sondes SOUIHI

Références

[1] Omar Adjali, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. Entity-aware cross-modal pretraining for knowledge-based visual question answering. In *ECIR*, pages 391–400, 2025.

[2] Omar Adjali, Paul Grimal, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. Explicit knowledge injection for knowledge-aware visual question answering about named entities. In *ACM ICMR*, pages 29–38, 2023.

[3] Omar Adjali, Paul Grimal, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. Multi-level information retrieval augmented generation for knowledge-based visual question answering. In *EMNLP*, pages 16567–16582, 2024.

[4] Robin Armingaud. GLiDRE : Modèle généraliste pour l'extraction de relations à l'échelle de documents. In *TextMine'25*, 2025.

[5] Robin Armingaud, Arthur Peuvot, Romaric Besançon, Olivier Ferret, Clément Maurer, Nasredine Semmar, and Sondes Souihi. CEA-LIST@EvalLLM2025 : des LLM mais sans modèle décodeur. In *EvalLLM*, 2025. 1re place au challenge.

[6] Robin Armingaud, Arthur Peuvot, Romaric Besançon, Olivier Ferret, Sondes Souihi, and Julien Tourille. CEA-List@EvalLLM2024 : prompter un très grand modèle de langue ou affiner un plus petit ? In *EvalLLM*, 2024. 1re place au challenge.

[7] Hugo Boulanger, Nicolas Hiebel, Olivier Ferret, Karën Fort, and Aurélie Névéol. Using Structured Health Information for Controlled Generation of Clinical Cases in French. In *Clinical NLP*, pages 172–184, 2024.

[8] Galo Castillo-López, Gael de Chalendar, and Nasredine Semmar. Intent recognition and out-of-scope detection using LLMs in multi-party conversations. In *SIGDial*, pages 504–512. Association for Computational Linguistics, 2025.

[9] Galo Castillo-López, Alexis Lombard, Nasredine Semmar, and Gaël de Chalendar. How ddair you ? disambiguated data augmentation for intent recognition. In *EACL*, 2026.

[10] Clément Cornet and Hervé Le Borgne. Deleuzian representation : Unsupervised interpretable concept extraction from neural networks. In *ICLR*, 2026. Oral presentation.

[11] Salim Abdou Daoura, Olivier Ferret, Romaric Besançon, and Sondes Souihi. Omnivent : Annotating a pretraining corpus for event extraction with a confidence-guided approach. In *AAAI 2026 Workshop on New Frontiers in Information Retrieval*, 2025.

[12] Perla Doubinsky, Nicolas Audebert, Michel Crucianu, and Hervé Le Borgne. Multi-attribute balanced sampling for disentangled GAN controls. *Pattern Recognition Letters*, 162 :56–62, 2022.

[13] Perla Doubinsky, Nicolas Audebert, Michel Crucianu, and Hervé Le Borgne. Semantic generative augmentations for few-shot counting. In *WACV*, 2024.

[14] Akram Elbouanani, Evan Dufraisse, and Adrian Popescu. Analyzing political bias in LLMs via target-oriented sentiment classification. In *ACL*, 2025.

[15] Akram Elbouanani, Evan Dufraisse, Aboubacar Tuo, and Adrian Popescu. CEA-LIST at CheckThat ! 2025 : Evaluating LLMs as detectors of bias and opinion in text. In *Working Notes of CLEF 2025*, 2025. 1re place en arabe et polonais.

[16] Akram Elbouanani, Adrian Popescu, and al. A scalable entity-based framework for auditing bias in LLMs, 2026. Prépublication arXiv.

[17] Alexandre Fournier-Montgieux, Adrian Popescu, Bertrand Luvison, and Hervé Le Borgne. Reliable and reproducible demographic inference for fairness in face analysis, 2025. Prépublication arXiv.

[18] Alexandre Fournier-Montgieux, Michaël Soumm, Adrian Popescu, Bertrand Luvison, and Hervé Le Borgne. Fairer analysis and demographically balanced face generation for fairer face verification. In *WACV*, 2025.

[19] Paul Grimal, Hervé Le Borgne, and Olivier Ferret. Unique step refinement for transformer-based generative models. In *ICPR*, 2026.

[20] Paul Grimal, Hervé Le Borgne, Olivier Ferret, and Julien Tourille. TIAM — A metric for evaluating alignment in text-to-image generation. In *WACV*, pages 2890–2899, 2024.

[21] Paul Grimal, Michaël Soumm, Hervé Le Borgne, Olivier Ferret, and Akihiro Sugimoto. SAGA : Learning signal-aligned distributions for improved text-to-image generation. In *AAAI*, 2026.

[22] Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. Can synthetic text help clinical named entity recognition ? a study of electronic health records in French. In *EACL*, pages 2320–2338, 2023.

[23] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G. Moreno, and Jesus Lovon. ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities. In *SIGIR*, 2022.

[24] Frédéric Lin, Biruk Abere Ambaw, Adrian Popescu, Hejer Ammar, Romaric Audigier, and Hervé Le Borgne. CaMiT : A time-aware car model dataset for classification and generation. In *NeurIPS*, 2025.

[25] Mehrad Moradshahi, Tianhao Shen, Kalika Bali, Monojit Choudhury, Gaël de Chalendar, Anmol Goel, Sungkyun Kim, Prashant Kodali, Ponnurangam Kumaraguru, Nasredine Semmar, Sina J. Semnani, Jiwon Seo, Vivek Seshadri, Manish Shrivastava, Michael Sun, Aditya Yadavalli, Chaobin You, Deyi Xiong, and Monica S. Lam. X-risawoz : High-quality end-to-end multilingual dialogue datasets and few-shot agents. In *ACL*, 2023.

[26] Achir Oukelmoun, Nasredine Semmar, Gaël de Chalendar, Clement Cormi, Mariame Oukelmoun, Eric Vibert, and Marc-Antoine Allard. Detecting omissions in LLM-generated medical summaries. In *EMNLP - Industry Track*, pages 325–337, 2025.

[27] Luis Palacios Medinacelli, Matteo Morelli, Gaël de Chalendar, Mykola Liashuha, Jaonary Rabarisoa, Lucas Labarussiat, and Raphaël Lallement. Bridging Natural Language Understanding, Symbolic Planning, and Robotic Execution through Hybrid AI Systems. In *Proceedings of the 2nd Workshop on Composite AI*, Montréal, Canada, August 2025.

[28] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. AVAE : adversarial variational auto encoder. *CoRR*, abs/2012.11551, 2020.

[29] Aboubacar Tuo, Romaric Besançon, Olivier Ferret, and Julien Tourille. Few-shot event argument extraction based on a meta-learning approach. In *NAACL*, pages 146–153, 2024.

AfIA
Association française
pour l'Intelligence Artificielle

GREYC/DIAMA∩T : Document, Interactions, Apprentissage Multimodal, Accessibilité, tal ∩ vision, Texte

GREYC UMR 6072 / DIAMA∩T
CNRS, Université de Caen Normandie
https://www.greyc.fr/equipes/diamant/

Fabrice MAUREL
Co-Responsable de l'équipe DIAMA∩T
fabrice.maurel@unicaen.fr

Laure LEJEUNE
Co-Responsable de l'équipe DIAMA∩T
laure.lejeune@unicaen.fr

Gaël DIAS
Directeur-adjoint du GREYC
gael.dias@unicaen.fr

Présentation générale

L'équipe DIAMA∩T (**D**ocument, **I**nteractions, **A**pprentissage **M**ultimodal, **A**ccessibilité, tal ∩ vision, **T**exte) du GREYC (UMR 6072) travaille sur la multimodalité texte / image / son / vidéo, articulée en une chaîne scientifique continue : produire des données multimodales situées, apprendre des modèles robustes, contrôlables et explicables, puis concevoir et évaluer des interactions en situation d'usage. L'équipe adopte une démarche *AI4Goods* et *AI4Health*, où les terrains applicatifs — santé mentale, accessibilité cognitive et non visuelle, sport et bien-être — servent de catalyseurs à des verrous méthodologiques transversaux : apprentissage multitâches et multivues, fusion cross-modale, alignement de représentations, apprentissage faiblement supervisé. Elle rassemble une quarantaine de personnes, permanents, associés et non permanents confondus, et porte ou co-porte quinze projets sur la période 2018–2030 (Horizon Europe MSCA-DN, France 2030, ANR, BPI, FHU, GIP, RIN/FEDER). Son activité s'organise en trois axes, chacun portant ses projets, ses ressources et ses plateformes.

Principaux thèmes de recherche

Axe 1 — Données multimodales

Cet axe porte sur l'acquisition, la structuration, l'annotation et l'anonymisation de corpus multimodaux « écologiques », collectés en situation réelle et souvent rares ou sensibles. Il constitue le socle empirique des deux autres axes.

L'équipe a produit ETR-fr, première ressource française de simplification entièrement conforme aux recommandations européennes « Facile à Lire et à Comprendre » (FALC) : 523 paires de paragraphes alignés issues d'ouvrages de littérature jeunesse adaptés, complétées d'un jeu de test hors domaine et d'une grille d'évaluation humaine de 36 items dérivée de ces mêmes recommandations [6]. Le projet ABC a donné lieu à MONET, corpus d'oculométrie et de mesures neuro-affectives recueilli au musée des Beaux-Arts de Caen auprès d'environ cent cinquante visiteurs, en observation individuelle puis en binôme, et complété par des questionnaires neuropsychologiques [8] ; il sert notamment à l'étude de l'ancrage visuel temporel au cours de la médiation artistique. Le projet INDYX traite le passage à l'échelle du Web : l'équipe y a construit un corpus multilingue de 1 689 pages de presse en cinq langues (anglais, chinois, grec, polonais, russe) pour évaluer les extracteurs de contenu principal et mettre en évidence leur biais anglocentré [3]. Sur le versant désinformation, Par-ITA fournit un corpus parallèle supervisé par des humains pour la neutralisation de titres hyperpartisans en italien [10], complété par le corpus trilingue PartisanLens.

Le projet PARTAGES (France 2030, communs numériques pour l'IA générative en santé) vise à démocratiser l'usage des grands modèles de langue auprès des professionnels de santé. L'équipe y est responsable de la génération de comptes rendus médicaux synthétiques : elle y développe des méthodes

de génération augmentée de recherche (RAG) enrichies de connaissances externes, ainsi qu'un cadre d'évaluation rigoureux combinant métriques automatiques supervisées et non supervisées, jugement par grand modèle de langue et annotation manuelle. Cinq mille comptes rendus synthétiques doivent être produits dans ce cadre. Le projet VigilanS-IA a permis de constituer et d'anonymiser un corpus clinique rare : le registre structuré des 5 954 patients suivis par le dispositif de recontact VigilanS en Normandie entre 2016 et 2023, complété des comptes rendus textuels d'appels de suivi pour 327 d'entre eux. EMERVA compare les productions FALC générées automatiquement à celles de collégiens plurilingues sur des textes à forte charge culturelle et figurative.

Axe 2 — Apprentissage profond multimodal

Cet axe développe des modèles capables d'aligner des représentations issues de modalités différentes, y compris sur de petits corpus spécialisés, et privilégie l'hybridation : guider les modèles génératifs par des contraintes linguistiques, discursives ou cognitives et par des métriques d'évaluation dédiées.

En génération FALC, l'équipe a montré qu'un apprentissage multitâches conjoint — génération FALC, simplification et résumé — améliore systématiquement les approches monotâches, que ce soit par génération augmentée de recherche ou par adaptation paramétrique de faible rang, la première généralisant mieux hors domaine et la seconde dominant en domaine [7]. Elle a également établi que de petits modèles pré-entraînés adaptés de façon parcimonieuse égalent des modèles de langue bien plus grands tout en généralisant mieux [6]. Sur la détection de contenus hyperpartisans, fallacieux ou nocifs, une comparaison systématique sur dix corpus et cinq langues montre que l'apprentissage en contexte reste en retrait d'un ajustement spécialisé, y compris sur des modèles de taille modeste [11]. L'analyse discursive complète ce dispositif : un score de généricité dérivé d'annotations clausales, croisé avec la théorie de la structure rhétorique, révèle que les textes générés sont plus génériques que les textes humains et que certains modèles conservent une cohérence bien structurée à forte généricité, de quoi accroître leur crédibilité apparente [4].

Le projet MENTAL.AI@CaeSAR (AMI Excellences du programme France 2030 CaeSAR) pose les bases d'une médecine dite 6P — personnalisée, prédictive, préventive, participative, fondée sur la preuve et intégrée au parcours de soin — pour la santé mentale et cérébrale, en mobilisant marqueurs numériques et modèles génératifs pour la détection précoce, le suivi et l'estimation du niveau de dépression. Une représentation par graphes multivues des entretiens, séparant les tours du patient et du clinicien, y atteint l'état de l'art sans recourir à des connaissances externes [1] ; une architecture fondée sur les symptômes et enrichie d'annotations d'experts fait apparaître une corrélation entre le comportement du modèle et les tendances des praticiens [2]. Sur le corpus VigilanS, la comparaison systématique de modèles tabulaires, textuels et de fusion établit qu'une fusion précoce devance toutes les approches unimodales, sans qu'aucune configuration n'atteigne encore une sensibilité et une spécificité simultanément suffisantes pour un usage clinique. L'explicabilité est également travaillée par la détection d'anomalies [5] et l'alignement intermodal par des représentations textuelles informées par le visuel [9]. Le projet ECOFALC développe des méthodes hybrides de génération FALC intégrant une *syntaxe émotionnelle*. Le réseau doctoral HYBRIDS (Horizon Europe MSCA-DN) construit une *intelligence hybride* qui injecte les connaissances structurées des sciences humaines et sociales dans les outils de TAL et les algorithmes d'apprentissage profond, afin de répondre à la désinformation — rumeurs, fausses nouvelles, discours de haine visant des minorités vulnérables — par une analyse fine du discours public.

Axe 3 — Interactions et évaluations multimodales

Cet axe relie modèles, dispositifs interactifs et mesures cognitives en situation réelle. L'objectif est d'intégrer l'humain dans la boucle d'apprentissage — jusqu'à un apprentissage par renforcement à partir de retours humains (*Reinforcement Learning from Human Feedback*, RLHF) enrichi par signaux physiologiques — et de mesurer engagement, attention et charge cognitive plutôt que de les inférer *a posteriori*.

L'accessibilité non visuelle en constitue le fil le plus ancien. La plateforme TagThunder transpose la sémantique morpho-dispositionnelle des pages Web en un nuage de tags auditif au moyen de synthèses vocales concurrentes spatialisées ; elle décline deux familles de stratégies, le survol et le balayage, de la vocalisation séquentielle à des parcours allocentriques fondés sur la similarité sémantique [13]. La plateforme TacTinet explore la voie complémentaire du retour tactile sur terminal mobile pour restituer la mise en page des documents [12].

Le projet STP-MEC vise à fonder scientifiquement la notion de texte FALC : un texte accessible ne se réduit pas à une simplification lexicale et syntaxique, il doit préserver conjointement ses charges linguistique, émotionnelle et cognitive, qui conditionnent l'attention, la motivation et la mémorisation du lecteur. Il articule des modèles génératifs multimodaux produisant un FALC cohérent aux plans syntaxique, sémantique et émotionnel, des métriques associant indicateurs linguistiques et marqueurs affectifs, et la validation empirique de la chaîne émotion → engagement → mémoire par protocoles cognitifs et neurophysiologiques, du FALC standard au FALC émotionnel puis multimodal texte/pictogramme.

La plateforme A2MIMO couple, au sein d'un agent conversationnel incarné, le suivi multimodal des marqueurs verbaux et non verbaux, la modélisation de l'alliance thérapeutique et la restitution clinique [14]. Les projets NEO-REEDUC, SIRENE et 2E2MR étendent ces méthodes d'évaluation instrumentée à la rééducation motrice et cognitive, au niveau d'éveil en conduite de navire et aux profils cognitifs experts en contexte sportif ; NEUROPSAI et M-PULSE en portent les résultats vers des dispositifs industriels.

Membres permanents

- Nicolas BENGUIGUI (PR)
- Youssef CHAHIR (MCF HDR)
- Gaël DIAS (PR)
- Laure LEJEUNE (MCF)
- Fabrice MAUREL (MCF HDR)
- François RIOULT (MCF HDR)
- Marc SPANIOL (PR)

Références

[1] Navneet Agarwal, Gaël Dias, and Sonia Dollfus. Multi-view graph-based interview representation to improve depression level estimation. *Brain Informatics*, 11(14), 2024.

[2] Navneet Agarwal, Kirill Milintsevich, Lucie Metivier, Maud Rotharmel, Gaël Dias, and Sonia Dollfus. Analyzing symptom-based depression level estimation through the prism of psychiatric expertise. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 974–983, Torino, Italia, May 2024. ELRA and ICCL.

[3] Aurélien Bournonville, Fabrice Maurel, François Rioult, Gaël Dias, Guillaume Pitel, Thomas Largillier, and Erwan Marchand. Multilingual evaluation of main content extractors for web pages. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025)*, Padua, Italy, 2025. hal-05040908.

[4] Søren Kirkegaard Fomsgaard, Martial Pastor, Gaël Dias, and Nelleke Oostdijk. Discourse realization of generics in human and LLM-generated texts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pages 32938–32960, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[5] Amine Youcef Khodja, Jérémie Pantin, Gaël Dias, and Fabrice Maurel. Anomaly detection using source-reversed flow matching. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP 2026)*, Budapest, Hungary, 2026.

[6] François Ledoyen, Gaël Dias, Alexis Lechervy, Jérémie Pantin, Fabrice Maurel, Youssef Chahir, Elisa Gouzonnat, Mélanie Berthelot, Stanislas Moravac, Armony Altinier, and Amy Khairalla. Inclusive easy-to-read generation for individuals with cognitive impairments. In *Proceedings of the 28th European Conference on*

Artificial Intelligence (ECAI 2025), Bologna, Italy, 2025. hal-05175510, arXiv :2510.00691.

[7] François Ledoyen, Gaël Dias, Jérémie Pantin, Alexis Lechervy, Fabrice Maurel, and Youssef Chahir. Facilitating cognitive accessibility with LLMs : A multi-task approach to easy-to-read text generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, pages 11771–11797, Suzhou, China, 2025. arXiv :2510.00662.

[8] Raphaëlle Lemaire, Azamat Kaibaldiyev, Eléonore Mariette, Débora Viglieri, Jérémie Pantin, Alexis Lechervy, Fabrice Maurel, Gaël Dias, Gaëtane Blaizot, Véronique Agin, Nicolas Poirel, Eric Bui, Hervé Platel, Denis Vivien, and Youssef Chahir. MONET : Museum observatory of neuro-affective eye-tracking. In *Proceedings of the 34th ACM International Conference on Multimedia (MM 2026)*, Rio de Janeiro, Brazil, 2026. À paraître.

[9] Raphaëlle Lemaire, Jérémie Pantin, Alexis Lechervy, Azamat Kaibaldiyev, Fabrice Maurel, Gaël Dias, and Youssef Chahir. Visually informed text representations for visual contextual classification of arts. In *Proceedings of the 28th International Conference on Pattern Recognition (ICPR 2026)*, Lyon, France, 2026. À paraître.

[10] Michele Joshua Maggini, Søren Fomsgaard, Michele Maestroni, Gaël Dias, and Pablo Gamallo. Par-ita : Benchmarking Seq2Seq and LLMs on a human-supervised parallel corpus for italian hyperpartisan neutralization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, San Diego, USA, 2026.

[11] Michele Joshua Maggini, Dhia-Eddine Merzougui, Rabiraj Bandyopadhyay, Gaël Dias, Fabrice Maurel, and Pablo Gamallo. Are LLMs enough for hyperpartisan, fake, polarized and harmful content detection ? evaluating in-context learning vs. fine-tuning. *Proceedings of the International AAAI Conference on Web and Social Media (ICSWM 2026)*, 20(1) :1551–1572, 2026. arXiv :2509.07768.

[12] Fabrice Maurel, Gaël Dias, Waseem Safi, Jean-Marc Routoure, and Pierre Beust. Layout transposition for non-visual navigation of web pages by tactile feedback on mobile devices. *Micromachines*, 11(4) :376, 2020. hal-02545507.

[13] Dhia-Eddine Merzougui, Nilesh Tete, Fabrice Maurel, Gaël Dias, Mohammed Hasanuzzaman, Aurélien Bournonville, Edgar Madelaine, Thomas Berthelin Le Tellier, François Ledoyen, Laure Poutrain-Lejeune, François Rioult, and Jérémie Pantin. Concurrent speech and auditory tag clouds for non-visual web interaction. In *Interspeech 2025*, Rotterdam, The Netherlands, 2025. hal-05110274.

[14] Soumaya Sabry, Antoine Cochard, and Gaël Dias. A2mimo : Agent conversationnel animé et analyse multimodale pour la santé mentale. Soumis au Workshop Affects, Compagnons Artificiels et Interactions (WACAI 2026), Paris, France, 2026.

AfIA
Association française
pour l'Intelligence Artificielle

Inalco/ERTIM : Équipe de Recherche Textes, Informatique, Multilinguisme

ERTIM
https://ertim.inalco.fr

Damien NOUVEL
Directeur
damien.nouvel@inalco.fr

Pierre MAGISTRY
Directeur adjoint
pierre.magistry@inalco.fr

Textes, Informatique, Multilinguisme

L'Equipe de Recherche Textes, Informatique, Multilinguisme (ERTIM) est une équipe de recherche, unité propre de l'Inalco, créée en 2005, qui travaille essentiellement en Traitement Automatique des Langues (TAL).

Le TAL, généralement considéré comme un domaine d'application de l'Intelligence Artificielle (IA) fait appel à l'informatique, aux mathématiques, à la linguistique. L'objectif principal est de concevoir ou d'exploiter des moyens (modèles et algorithmes) pour analyser ou générer des données en langues, selon une modalité écrite ou orale. Les tâches spécifiques sont de diverses natures : acquisition et numérisation de données textuelles, modélisations linguistiques ou sémantiques, extraction d'information, fouille de données, traduction automatique, système de dialogue, etc.

L'ERTIM s'intéresse plus particulièrement aux questions méthodologiques (collecte de données, numérisation, annotation, évaluation qualitative ou quantitative) dans des contextes fortement multilingues, avec un intérêt marqué pour les langues peu dotées du domaine de l'Inalco. Les activités de l'équipe sont déclinées selon quatre axes : humanités numériques, diversité des langues, méthodologie du TAL, acquisition de l'information linguistique.

Activités et projets

Les activités de l'équipe s'articulent autour des axes présentés ci-dessous et se matérialisent par des projets de recherche. Les méthodes et outils utilisés évoluent rapidement selon les avancées en traitement automatique des langues.

Axe : TAL et sécurité cognitive (resp. : Mathieu Valette)

Les travaux de l'axe explorent l'intersection critique entre l'IA, le TAL et la sécurité cognitive. Ils visent à détecter, comprendre, mesurer et contrer les risques d'influence et de manipulation en contexte numérique (réseaux sociaux) et issues de systèmes d'IA (modèles de langage, agents conversationnels). Pour ce faire, ils mobilisent des partenariats pluridisciplinaires (sciences cognitives, psychologie, géopolitique, information-communication) et des méthodologies expérimentales (simulations, tests en environnement réel, ou évaluation en boîte noire des LLMs). L'axe accueille actuellement une thèse et trois projets de recherche :

Le projet GECKO (Laboratoire de conception pour la guerre cognitive) aborde la guerre cognitive via une approche pluridisciplinaire (sciences cognitives, communication, politique, études aréales). Son objectif est d'analyser les manipulations intentionnelles de la cognition individuelle ou collective (raisonnement, prise de décision) dans des contextes de crise fictive. Le projet vise à préparer les acteurs civils et militaires (décisionnels ou opérationnels) aux risques liés à la sécurité nationale en développant des dispositifs d'exploration pour anticiper et contrer ces menaces (AID ASTRID 2023-2026 – Inalco, IMS Bordeaux, IRSEM).

Le projet IACARI (Intelligence Artificielle Conversationnelle Anticipative pour les Récits d'Influence) se concentre sur l'évaluation des capacités d'influence des IA conversationnelles dans des environnements réels (réseaux sociaux). Il propose de prototyper un dispositif pour tester, à grande

échelle, l'impact d'une IA intervenant de manière transparente dans des discussions : apport d'informations, renforcement de récits positifs, ou inflexion des débats. L'enjeu est de mesurer l'efficacité de ces interventions sur la résilience cognitive et informationnelle, notamment face à la post-vérité (science, droit, relations internationales). Le projet rassemble des experts en géopolitique, IA, TAL, psychologie et interaction humain-machine (AID ASTRID 2026-2028 – Inalco, IMS Bordeaux, IFG Lab)

Le projet Eval-IA (Évaluation en boîte noire des capacités d'influence et de manipulation des modèles d'IA) propose une méthodologie d'évaluation en « boîte noire » pour analyser les capacités d'influence des LLMs. Il combine le fuzzing (tests dynamiques des entrées/sorties), l'analyse sémantique automatisée et des comparaisons à grande échelle entre les réponses des modèles et des données humaines. L'objectif est de révéler les biais culturels et normatifs des modèles, qui constituent une influence implicite pouvant affecter les perceptions et décisions des utilisateurs (INESIA 2026-2032 – Inria, Inalco).

Axe : diversité des langues (resp. : Pierre Magistry et Ilaine Wang)

En cohérence avec la mission de conservation des langues du monde portée par l'Inalco, l'ERTIM s'intéresse à la question de la diversité des langues dans les techniques de traitement automatique. L'accent est mis sur les langues dites « peu dotées » et les langues d'héritage. Deux angles d'approche sont envisagés : d'une part, étudier ce que ces langues posent comme questions spécifiques pour le TAL et pour la nécessaire adaptation des méthodes (corpus et autres ressources limités, effets de l'apprentissage par transfert, absence de standardisation...) ; d'autre part, montrer ce que le TAL peut apporter dans la description et la préservation de ces langues. Une langue sans outils de traitement automatique ou mal prise en charge par les systèmes informatiques est aujourd'hui d'autant plus menacée.

ERTIM s'intéresse au cas des langues sinitiques dans le cadre du projet DLS-HN (Diversité des Langues Sinitiques et Humanités Numériques") et en abordant la question sous l'angle de la variation temporelle (diachronique), géographique (dialectale/diatopique) mais aussi grapholinguistique (rapport langue-écriture). Pour ces trois axes, nos travaux portent d'une part sur l'évaluation et la conception des méthodes de TAL sur des données situées à différentes positions le long de ces axes, et d'autre part sur la diffusion de ces méthodes et leurs applications. Pour l'axe géographique, nous avons d'abord travaillé à la constitution de ressources à la fois écrites et orales à l'aide de méthodes automatiques. Nous avons ainsi constitué un treebank parallèle permettant de comparer le taigi, parlé à Taïwan, à la fois à une langue sinitique distante (le mandarin) et à une langue sinitique proche (le teochew) [4], à l'aide de méthodes de transfert appliquées au parsing. Côté oral, nous avons développé un outil de reconnaissance automatique de la parole (*speech-to-text*), permettant transcrire les données orales que nous avons recueillies et de faire d'y faire des requêtes orales [2]. Un système de synthèse vocale pour le teochew (*text-to-speech*) a également été construit afin de compléter un dictionnaire teochew existant [6, 7]. Ces travaux ont par ailleurs été réalisés avec une approche participative, en collaboration avec la communauté de locuteurs de teochew de France [5].

Axe : méthodologie du TAL (resp. : Damien Nouvel)

Les travaux scientifiques menés à l'ERTIM font majoritairement appel à des méthodes de TAL qui nécessitent une expertise des langues et corpus traités pour mieux concevoir, évaluer et exploiter les tâches réalisées. Les progrès technologiques en traitement automatique de données sont indéniables : ces dernières années, la mise au point et l'exploitation de méthodes relevant de l'IA et s'appuyant sur des réseaux de neurones pour construire des modèles de langues sont devenus la question centrale du TAL, autant dans l'industrie que dans la communauté scientifique. L'unité travaillera sur tous les aspects de ces technologies pour produire des études, concevoir des algorithmes ou des ressources.

Le projet DALiH (Digitizing Armenian Linguistic Heritage) vise à construire une plate-forme linguistique numérique inédite, unifiée, en libre accès et

open-source couvrant l'ensemble du spectre variationnel de la langue arménienne. Dans le cadre de ce projet, ERTIM est chargé des recherches portant sur les données orales, avec notamment un travail autour de l'identification automatique des variétés d'arménien, du calcul de la distance entre variétés, mais aussi de la création d'outils tels qu'un aligneur automatique et un outil de reconnaissance de la parole fonctionnel prenant en compte les variétés parlées. Ces travaux sont réalisés en collaboration avec des chercheurs des l'Inalco et en Arménie spécialistes de ces langues ou des technologies mobilisées [1, 3].

Axe : acquisition de l'information linguistique (resp. : Kata Gabor)

Les travaux de cet axe vont s'inscrire dans le domaine de la linguistique informatique. L'unité veillera à identifier, modéliser, extraire des informations pertinentes à la compétence linguistique à partir de données de performance, qui peuvent provenir de corpus ou, plus récemment, des représentations créées par des modèles de langues à partir de corpus. Dans ce contexte, la problématique de la distinction entre généralisation (statistique, linguistique) et la mémorisation lexicale présente un intérêt particulier. ERTIM cherchera à mettre en œuvre des méthodologies qui permettent de distiller des schémas linguistiques à partir de données, et de vérifier leur validité, en portant un regard critique sur les approches purement expérimentales (les « probes » par exemple) et une attention toute particulière à formuler des hypothèses linguistiques qui facilitent l'interprétation des expériences.

Chaire junior : Intelligence Artificielle pour les langues rares ou peu dotées (resp. : Valentina Fedchenko)

Le projet principal de la chaire porte sur le développement de méthodes pour faciliter et accélérer la documentation des langues. L'un des objectifs principaux est d'automatiser les différentes étapes de traitement des données issues du terrain linguistique, notamment la transcription, l'annotation et l'élaboration de ressources exploitables à partir d'enregistrements audio. Une attention particulière est accordée aux données brutes provenant des enquêtes de terrain et des archives sonores, qui constituent souvent des ressources importantes mais difficiles à exploiter en raison de leur volume, de leur hétérogénéité et de l'absence de transcriptions. Les méthodes développées doivent prendre en compte les caractéristiques propres aux langues peu dotées : faible disponibilité de données annotées, absence ou faible degré de standardisation orthographique, forte variation dialectale, diversité des pratiques de transcription et, dans certains cas, phénomènes de contact et de variation linguistique.

Un axe central de la chaire concerne la reconnaissance automatique de la parole et, plus particulièrement, l'adaptation de modèles acoustiques multilingues à des langues et variétés disposant de très peu de données annotées. L'approche privilégiée repose sur l'exploitation de la similarité des données afin d'identifier les langues ou variétés susceptibles de servir de sources pour l'apprentissage par transfert. Cette approche conduit à une réflexion plus générale sur la notion de similarité dans les modèles acoustiques.

Le projet CLingS vise à améliorer l'accès aux connaissances scientifiques produites dans des langues encore peu représentées dans les infrastructures numériques et les technologies du langage. Le projet porte sur sept langues : biélorusse, estonien, pendjabi, slovaque, taïwanais (Taigí), ukrainien et yiddish. Il prévoit l'extraction d'information à partir de corpus scientifiques dans plusieurs domaines, notamment la linguistique et la philologie, la médecine, les mathématiques, la géographie et le droit. Des méthodes de recherche d'information multilingue fondées notamment sur des approches de Graph-RAG seront développées afin d'identifier des documents et des passages pertinents dans différentes langues. Le projet prévoit également la mise au point d'outils d'alignement terminologique prenant en compte les correspondances entre langues, les différentes traditions scientifiques ainsi que les spécificités culturelles dans l'utilisation des termes, notamment dans le domaine des sciences humaines et sociales.

Le principal résultat attendu est une plateforme de recherche d'information scientifique multilingue et translingue, permettant d'effectuer des requêtes en langage naturel et d'accéder à des contenus

scientifiques dans plusieurs langues, avec des réponses fondées sur les sources et accompagnées de références bibliographiques précises.

D'une durée de trois ans, de janvier 2026 à décembre 2028, le projet est coordonné par l'Institut national des langues et civilisations orientales (Inalco). Le consortium réunit également l'Institute of Informatics de la Slovak Academy of Sciences (Bratislava), la Constantine the Philosopher University in Nitra et la National Academy for Educational Research (Taipei).

Membres permanents de l'équipe

- Bénédicte DIOT-PARVAZ AHMAD (MCF)
- Valentina FEDCHENKO (CPJ)
- Kata GABOR (MCF)
- Pierre MAGISTRY (MCF)
- Damien NOUVEL (MCF)
- Polina MIKHEL (MCF)
- Mathieu VALETTE (PR)
- Ilaine WANG (IGR)

Références

[1] Malajyan Arthur, Victoria Khurshudyan, Karen Avetisyan, Hossep Dolatian, and Damien Nouvel. Bi-dialectal ASR of Armenian from naturalistic and read speech. In Maite Melero, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 227–236, Torino, Italia, May 2024. ELRA and ICCL.

[2] Siman Chen, Ilaine Wang, Maxime Fily, and Pierre Magistry. Leveraging Speech Models for Audio-based Lexical Retrieval in Dictionaries : the Case of the Teochew Language. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference – SPEAKABLE workshop*, Language Resources and Evaluation Conference, Palma De Majorque, Spain, May 2026. ELRA Language Resources Association.

[3] Victoria Khurshudyan, Nadi Tomeh, Damien Nouvel, Anaïd Donabédian, and Chahan Vidal-Gorène. Proceedings of the lrec 2022 workshop on processing language variation : Digital armenian (digitam). 2022.

[4] Pierre Magistry and Ilaine Wang. Design of a mSUD Taiwan Taigi TreebankAligned on Mandarin and Teochew Translations. In *Proceedings of The 36th Conference on Computational Linguistics and Speech Processing (ROCLING 2024)*, Taipei, Taiwan, November 2024. ACLCLP.

[5] Pierre Magistry and Ilaine Wang. La science participative et l'ANR DiLSi. In Mathilde Hutin and Alice Millour, editors, *Actes de l'atelier Science Participative pour les Données et Corpus Linguistiques 2025 (ParCol)*, pages 3–5, Marseille, France, 2025. ATALA
ARIA.

[6] Pierre Magistry, Ilaine Wang, and Ty Eng Lim. Experiments on Speech Synthesis for Teochew, Can Taiwanese Help ? In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italy, May 2024.

[7] Agathe Wallet, Ilaine Wang, Emmett Strickland, and Pierre Magistry. Evaluating Speech Synthesis in a Nonstandardized, Multidialectal Context : A Teochew Case Study. In *Proceedings of the 13th edition of the Speech Synthesis Workshop*, Proceedings of the 13th edition of the Speech Synthesis Workshop, Leeuwarden, Netherlands, August 2025.

AfIA
Association française
pour l'Intelligence Artificielle

IRISA/ARCHIMÉDIA

IRISA UMR 6074
CNRS et Université de Rennes
https://team.inria.fr/archimedia/

Laurent AMSALEG
laurent.amsaleg@cnrs.fr

Introduction

De vastes collections de contenu multimédia numérique sont détenues par diverses institutions, organisations et entreprises, par exemple les chaînes de télévision, les stations de radio, les journaux. Ces collections, façonnées par des politiques éditoriales, contiennent une grande variété de types de documents et sont riches en contenu et en relations. L'utilisation de grands modèles de langage profonds et multimodaux a permis d'extraire des connaissances de haute qualité à partir de ces collections. Cependant, de nombreux problèmes de recherche restent ouverts, en particulier dans des contextes professionnels à enjeux élevés où des résultats précis et fiables sont essentiels.

Le projet ARCHIMÉDIA vise à étudier les problèmes scientifiques soulevés par l'utilisation d'ordinateurs pour analyser automatiquement ces collections. La plupart des problèmes que nous rencontrons sont liés à l'impact des propriétés et de l'usage de vastes collections de matériel multimédia dans des contextes professionnels sur les algorithmes. Parmi les principaux problèmes identifiés, on trouve :

- analyse ciblée : les professionnels ont besoin de réponses précises issues des collections, contrairement aux systèmes d'IA à usage général qui fournissent des résultats généralistes,
- collections dynamiques : les collections évoluent en permanence, nécessitant un apprentissage continu et des mises à jour, à la différence des systèmes d'IA statiques,
- efforts à long terme : les collections sont le fruit de projets à long terme avec des annotations et des métadonnées qui évoluent, ce qui limite l'efficacité des systèmes d'IA,
- contextes temporels : l'analyse des collections doit prendre en compte le contexte temporel des documents, car leur sens et leur pertinence peuvent changer avec le temps,
- analyse collaborative : les collections permettent un travail collaboratif sur le long terme entre plusieurs personnes, que les modèles actuels ne facilitent pas,
- ressources contraintes : de nombreuses solutions d'IA sont gourmandes en ressources et restent inaccessibles à la plupart des organisations.

ARCHIMÉDIA est une équipe conjointe avec la société Ouest-France. Le contexte professionnel privilégié est donc celui du journalisme, compte tenu de notre collaboration de longue date.

La recherche dans ARCHIMÉDIA

ARCHIMÉDIA est un projet de recherche conjoint académique-industriel visant à analyser de vastes collections multimédias dans des contextes réels. Il réunit des chercheurs universitaires et des chercheurs de l'entreprise Ouest-France.

Le projet se concentre sur l'étude des propriétés et de l'usage des collections multimédias en contexte professionnel, en s'appuyant notamment sur les archives uniques d'Ouest-France, qui comptent des millions d'articles, de photos, de pages, de vidéos et de podcasts, couvrant plus de 100 ans de matériel numérisé.

Le partenariat académique-industriel d'ARCHIMÉDIA est unique du fait de sa collaboration de longue durée (près de 30 ans), de son focus de recherche *in vivo* sur les environnements professionnels concrets, et de son insistance sur la résolution de problématiques réelles.

Il favorise les transferts de connaissances et de technologies, développe des prototypes appliqués de preuve de concept, et combine l'expertise académique en IA avec l'expérience industrielle de gestion de vastes archives multimédias.

Ce partenariat vise à créer des modèles neuronaux puissants capables de saisir l'information

contextuelle, en répondant à des défis tels que l'annotation, la fusion de connaissances et la prise en compte de la dimension temporelle de l'information. Ces modèles doivent être adaptés aux contextes professionnels, en mettant l'accent sur la durabilité, la confiance et l'explicabilité.

ARCHIMÉDIA explore deux axes de recherche principaux :

- Conception de modèles puissants : cet axe porte sur le développement de modèles neuronaux capables de traiter séparément ou conjointement diverses modalités (images, vidéos, audio, texte). L'objectif est d'améliorer les fondations de ces modèles en répondant à leurs besoins de supervision, à leur adaptabilité aux données évolutives, à leur précision et à leur capacité à générer des représentations discriminantes à haute dimension malgré la malédiction de la dimensionnalité. Les enjeux majeurs comprennent l'alignement et l'intégration des données provenant de modalités différentes, la garantie de la généralisation et de la robustesse, ainsi que le maintien de l'évolutivité.
- Couplage modèles et connaissances : cet axe vise à intégrer les grands modèles de langage (LLM) avec des bases de connaissances (KB) pour enrichir l'analyse de grandes collections multimédias. La recherche explore comment les LLM peuvent améliorer et compléter les KB, comment les KB peuvent être incorporées aux LLM lors du préentraînement, et comment un couplage étroit LLM KB permet de combiner leurs forces. Cela inclut la prise en compte des modèles multimodaux et des KB, des modèles et KB sensibles au facteur temporel, ainsi que la fusion de connaissances provenant de multiples sources.

Membres permanents

- Laurent AMSALEG (DR)
- Nicolas BÉCHET (MCF)
- Caio CORRO (MCF)
- Guillaume GRAVIER (DR)
- Damien LOLIVE (PR)
- Pascale SÉBILLOT (PR)
- Vincent SEGONNE (MCF)

AfIA
Association française
pour l'Intelligence Artificielle

IRIT/MELODI : MEthodes et ingénierie des Langues, des Ontologies et du DIscours

IRIT UMR 5505
CNRS, Toulouse INP, UT, UT2
https://www.irit.fr/departement/intelligence-artificielle/equipe-melodi/

Véronique MORICEAU
veronique.moriceau@irit.fr

Nathalie HERNANDEZ
nathalie.hernandez@irit.fr

Présentation générale

L'équipe MEthodes et ingénierie des Langues, des Ontologies et du DIscours (MELODI) se préoccupe de recherches articulant l'analyse et la formalisation de la langue naturelle d'une part, la représentation et la modélisation des connaissances d'autre part. De plus, l'équipe offre un éventail d'approches de la recherche allant d'études théoriques sur les éléments contribuant à la sémantique et à sa représentation formelle, sur les interactions et les structures de discours, à des études plus expérimentales basées sur des développements logiciels, la constitution de ressources (corpus, corpus annotés, lexiques et ontologies) et des évaluations en corpus. Une partie de ses travaux se situe dans une démarche d'ingénierie des connaissances pour outiller la construction d'applications faisant appel à ces caractérisations sémantiques et à des modèles de connaissances.

Pour le volet *langue naturelle*, les thématiques abordées par MELODI concernent :

- l'analyse et la description conceptuelle ou formelle de phénomènes linguistiques au niveau de la sémantique lexicale, de la phrase, du discours ou du texte ;
- la classification de texte, l'apprentissage automatique de représentations structurées pour le TAL.

Pour le volet *représentation des connaissances*, MELODI traite les points suivants :

- des méthodes et des principes de construction d'ontologies et de gestion de leurs évolutions ;
- la modélisation et la représentation de connaissances à l'aide de graphes ou de la logique.

À l'intersection de la langue naturelle et des représentations des connaissances MELODI aborde :

- la construction d'ontologies à partir de textes ;
- la caractérisation linguistique de relations sémantiques ;
- les méthodes de développement d'applications faisant appel à la dualité langue/connaissances.

Principaux thèmes de recherche

Langage

Nos travaux se concentrent sur les approches hybrides de traitement automatique des langues (TAL) pour l'analyse des textes et des conversations. L'objectif est de combiner les connaissances linguistiques (au niveau de la phrase et du discours) et extra-linguistiques pour inférer le contexte nécessaire à l'interprétation d'énoncés. Ces approches requièrent à la fois :

1. le développement de modèles d'apprentissage opérant sur l'interface sémantique/pragmatique pour l'extraction d'information, en particulier les expressions/structures implicites du langage,
2. l'explicabilité et interprétabilité de ces modèles,
3. la prise en compte du goulot d'étranglement lié au manque de données d'apprentissage ainsi qu'à leur qualité, souvent bruitée (e.g., données de chat ou réseaux sociaux).

Sur chacun de ces trois axes, nous proposons de nouvelles ressources linguistiques/modèles pour le français, mais aussi multilingues (anglais, italien, espagnol, arabe, etc.) [6, 8, 12, 18]. Nous explorons également des approches multimodales pour pallier le manque de contexte linguistique des énoncés, comme l'audio/parole [16] ou les interactions situées [19]. Les modèles sont évalués sur diverses tâches allant de la classification (faiblement) supervisée de textes à la génération pour des applications à fort impact social, comme la modération de contenu sur les réseaux sociaux [9], les biais politiques [10], la sécurité des biens et des personnes

[15], ou encore le domaine médical avec l'analyse de spécificités linguistiques liées à des addictions ou des troubles cognitifs [14, 17].

Avec l'avènement des grands modèles de langues pré-entrainés (LLMs), des problèmes de recherche complexes restent ouverts sur le plan théorique (biais, capacités de raisonnement, performances à appréhender la complexité) et éthique. Dans ce contexte, nous menons des travaux sur les hallucinations de ces modèles [2, 3], le multilinguisme, les aspects culturels, le résumé automatique ainsi que les stratégies de prompting/fine-tuning des LLMs pour le peuplement de graphes de connaissances [13].

Représentation des connaissances

Nos recherches étudient les ontologies et les graphes de connaissances selon deux axes. D'une part des problématiques d'ingénierie visent à définir des logiciels, modèles ou méthodes pour représenter des données spécifiques. D'autres travaux s'intéressent aux fondements philosophiques et logiques des représentations ontologiques dans des contextes appliqués [20, 5]. Ces deux axes complémentaires s'articulent autour des questions de nature des contenus et de construction, peuplement, alignement des ontologies et graphes de connaissances. Plusieurs approches de construction de modèles de connaissances sont étudiées :

- le peuplement de graphes à partir de données issues de sources hétérogènes, en particulier des données géolocalisées [4, 7]. La mise à disposition de ces données sous forme de graphes permet de les interroger par des requêtes élaborées mais accessibles aux utilisateurs finaux,
- l'enrichissement de graphes à partir d'informations extraites de textes par des approches d'apprentissage supervisé ou non supervisé avec de grands modèles de langue [11, 1].

Membres Permanents

- Nathalie AUSSENAC-GILLES (DR)
- Nicholas ASHER (DR)
- Adrien BARTON (CR)
- Farah BENAMARA (PR)
- Chloé BRAUD (CR)
- Nathalie HERNANDEZ (PR)
- Shufan JIANG (MCF)
- Véronique MORICEAU (MCF HDR)
- Philippe MULLER (PR)
- Laure VIEU (DR)

Références

[1] Amina Annane, Nathalie Aussenac-Gilles, and Mouna Kamel. BBO : BPMN 2.0 Based Ontology for Business Process Representation. In *Proceedings 20th European Conference on Knowledge Management (ECKM 2019)*, volume 1, pages 49–59, Lisbonne, Portugal, September 2019.

[2] Nicholas Asher and Swarnadeep Bhar. Strong hallucinations from negation and how to fix them. In Association for Computational Linguistics, editor, *Proceedings 62nd Annual Meeting of the Association forComputational Linguistics (ACL 2024)*, pages 12670–12687, Bangkok / Thailand, Thailand, August 2024. Association for Computational Linguistics.

[3] Nicholas Asher, Swarnadeep Bhar, Akshay Chaturvedi, Julie Hunter, and Soumya Paul. Limits for learning with language models. *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 236–248, September 2023.

[4] Nathalie Aussenac-Gilles, Catherine Comparot, Antoine Dupuy, Nabil El Malki, Ronan Tournier, Ba-Huy Tran, and Cassia Trojahn dos Santos. eCISE-OWL : Représentation OWL du Schéma de Données de l'Environnement Commun d'Échange d'Information pour le Domaine Maritime. In Fatiha Saïs, editor, *Actes IC 2022 - Journées d'ingénierie des Connaissances @ PFIA 2022*, Journées Francophones D'Ingénierie des Connaissances, pages 82–91, Saint-Etienne, France, June 2022. AFIA.

[5] Adrien Barton, Fumiaki Toyoshima, Laure Vieu, Paul Fabry, and Jean-François Ethier. The Mereological Structure of Informational Entities. In *Formal Ontology in Information Systems. Proceedings of the 11th International Conference (FOIS 2020)*, volume Volume 330 : Formal Ontology in Information Systems

of *Frontiers in Artificial Intelligence and Applications*, pages 201–215. IOS Press, 2020.

[6] Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang S. Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads. In *Findings of the Association for Computational Linguistics : EACL 2023*, volume 2023.findings-eacl, pages 686–696, DubrovniK, Croatia, May 2023. ACL : Association for Computational Linguistics, Association for Computational Linguistics.

[7] William Charles, Nathalie Aussenac-Gilles, and Nathalie Jane Hernandez. Diachronical geometry without polygons : the extended HHT ontology for heterogeneous geometrical representations. In *Lecture Notes in Computer Science (LNCS)*, volume 15233 of *Lecture Notes in Computer Science*, pages 80–96, Baltimore (MA), United States, November 2024. Springer Nature Switzerland.

[8] Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. An Annotated Corpus for Sexism Detection in French Tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference 2020*, pages 1–7, online, France, May 2020. ELRA : European Language Resources Association.

[9] Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. Emotionally Informed Hate Speech Detection : A Multi-target Perspective. *Cognitive Computation*, 14 :322–352, June 2021.

[10] Nicolas Devatine, Philippe Muller, and Chloé Braud. An Integrated Approach for Political Bias Prediction and Explanation Based on Discursive Structure. In *Findings of the Association for Computational Linguistics (EACL 2023)*, pages 11196–11211, Toronto, Canada, July 2023. ACL : Association for Computational Linguistics.

[11] Mohamed Ettaleb, Véronique Moriceau, Mouna Kamel, and Nathalie Aussenac-Gilles. The contribution of LLMs to relation extraction in the economic field. In *Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal) - Proceedings of the workshops*, pages 175–183, Abu Dhabi, United Arab Emirates, January 2025. Chung-Chi Chen and Antonio Moreno-Sandoval and Qianqian Xie and Jimin Huang and Sophia Ananiadou and Hsin-Hsi Chen, Association of COmputational Linguistics.

[12] Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. IDAT at FIRE2019 : Overview of the Track on Irony Detection in Arabic Tweets. In *11th Forum for Information Retrieval Evaluation (FIRE)*, Kolkata, India, December 2019.

[13] Khorashadizadeh Hanieh, Amara Fatima Zahra, Ezzabady Morteza, Ieng Frédéric, Tiwari Sanju, Mihindukulasooriya Nandana, Groppe Jinghua, Sahri Soror, Farah Benamara, and Groppe Sven. Research Trends for the Interplay between Large Language Models and Knowledge Graphs. In *1st International Workshop on Data Management Opportunities in Unifying Large Language Models + Knowledge Graph. Workshop at the 50th International Conference on Very Large Data Bases (VLDB 2024)*, Guangzhou, China, August 2024.

[14] Jeanne Jerpan, Véronique Moriceau, Alexandrine Salis, Remy Klein, François Olivier, and Juliette Salles. Changes in suicide-related tweets before and during the COVID-19 pandemic in France : The importance of social media monitoring in public health prediction. *L'Encéphale*, 50(5) :516–523, October 2024.

[15] Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane. A three-level classification of French tweets in ecological crises. *Information Pro-*

cessing and Management, 57(5) :1–46, September 2020. Pre-print.

[16] Adrien Lafore, Clément Pagés, Leila Moudjari, Sebastião Quintas, Hervé Bredin, Thomas Pellegrini, Farah Benamara, Isabelle Ferrané, Jérôme Bertrand, Marie-Françoise Bertrand, Véronique Moriceau, and Jérôme Farinas. IRIT-MFU Multi-modal systems for emotion classification for Odyssey 2024 challenge. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2024)*, Proc. The Speaker and Language Recognition Workshop (Odyssey 2024), pages 296–302, Québec, Canada, June 2024. ISCA.

[17] Chuyuan Li, Maxime Amblard, Chloé Braud, Caroline Demily, Nicolas Franck, and Michel Musiol. Investigating non lexical markers of the language of schizophrenia in spontaneous conversations. In Chloé Braud, Christian Hardmeier, Junyi Jessy Li, Annie Louis, Michael Strube, and Amir Zeldes, editors, *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 20–28, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[18] Romain Meunier, Farah Benamara, Véronique Moriceau, Zhongzheng Qiao, and Savitha Ramasamy. CrisisTS : Coupling Social Media Textual Data and Meteorological Time Series for Urgency Classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 16082–16099, Vienna, France, July 2025. Association for Computational Linguistics.

[19] Maike Paetzel-Prüsmann, Julie Hunter, Kranti Chalamalasetti, Kate Thompson, Alexandros Nicolaou, Ozan Güngör, David Schlangen, and Nicholas Asher. Conversational Programming for Collaborative Robots. In *en ligne*, pages 1–5, Philadelphia, United States, May 2022. 39th edition of the IEEE International Conference on Robotics and Automation (ICRA 2022).

[20] Cédric Tarbouriech, Laure Vieu, Adrien Barton, and Jean-François Ethier. From slot mereology to a mereology of slots. *Applied Ontology*, 19(2) :181–230, 2024.

AfIA
Association française
pour l'Intelligence Artificielle

LabHC : Laboratoire Hubert Curien

Laboratoire Hubert Curien UMR 5516
Université Jean Monnet Saint-Etienne, CNRS,
Institut d'Optique Graduate School
https://laboratoirehubertcurien.univ-st-etienne.fr/

Antoine GOURRU
antoine.gourru@univ-st-etienne.fr

Présentation générale

Le Laboratoire Hubert Curien est une unité mixte de recherche associant le CNRS, l'Université Jean Monnet et l'Institut d'Optique Graduate School. Ses activités couvrent un large spectre de thématiques centrées autour de l'informatique et de la physique, notamment les systèmes embarqués, la sécurité, la photonique, et l'apprentissage automatique.

L'équipe **Machine Learning** est structurée en deux sous-équipes projet et une équipe Inria transversale. Elle s'organise autour des équipes projet suivantes :

- Malice (Inria - Machine Learning for Physics)
- Foundations and Theory of Machine Learning
- Responsible and Trustworthy Machine Learning

Le thème *Foundations and Theory of Machine Learning* vise à établir des fondements théoriques pour l'apprentissage statistique, en étudiant notamment la généralisation, l'optimisation, le transport optimal et l'apprentissage de représentations. Le thème *Responsible and Trustworthy Machine Learning* porte sur la frugalité, l'explicabilité et l'équité des modèles, avec un accent sur la détection et la réduction des biais.

Une grande partie des membres de l'équipe concentre ses recherches sur le **traitement automatique du langage** (TAL). Ces travaux portent sur la modélisation et l'analyse de données textuelles à l'aide de méthodes d'apprentissage automatique et profond, notamment pour l'apprentissage de représentations sémantiques, les approches frugales et équitables, la production de ressources textuelles originales pour la communauté et l'intégration de connaissances externes, en particulier sous forme de graphes de connaissances.

1. https://www.anr-dike.fr/

Principaux thèmes de recherche

Mesure et réduction des biais

Les travaux de l'équipe proposent une approche intégrée de l'équité en TAL, combinant ressources d'évaluation, analyse empirique et méthodes de débiaisage. Une contribution centrale est la conception d'un jeu de données en français dédié à l'évaluation de dimensions éthiques, annoté pour l'alignement moral, qui met en évidence la sensibilité des modèles aux variations de contexte culturel [12]. Sur le plan méthodologique, l'équipe introduit des approches de classification équitable fondées sur des contraintes d'indépendance aux attributs sensibles, notamment via une formulation basée sur l'information mutuelle dite de Wasserstein. Cette approche permet de découpler les représentations liées à la tâche et aux attributs sensibles sans nécessiter d'annotations explicites, tout en préservant les performances [10]. Elle est accompagnée de garanties théoriques fortes sur l'amélioration de l'équité [11]. En complément, plusieurs travaux de l'équipe analysent les biais dans les grands modèles de langue pour la détection de positions. Nous avons notamment montré que ces modèles exploitent des stéréotypes implicites pour inférer la position exprimée dans un texte sur un sujet donné, révélant des corrélations indésirables entre contenu linguistique et attributs sociodémographiques [1]. Cet axe de travail s'inscrit principalement dans le cadre de l'ANR Diké portée par le laboratoire [1].

TAL frugal

L'équipe développe des approches de TAL frugales visant à optimiser le compromis entre performance et coût computationnel. Ses travaux ré-

cents se concentrent sur des architectures multi-modèles dans lesquelles un mécanisme de *routage* sélectionne dynamiquement, pour chaque requête, le modèle le plus adapté tout en minimisant le coût [14]. Cette stratégie permet de mobiliser des modèles légers pour les cas simples et de recourir à des modèles plus coûteux uniquement lorsque nécessaire. L'équipe a notamment produit un travail de synthèse structurant ce domaine, proposant une taxonomie des stratégies de routage et analysant les compromis entre qualité, coût et ressources [15].

Questions-réponses et graphes de connaissances

Les travaux de l'équipe sur le *question answering* visent à interroger des données complexes en langage naturel en s'appuyant sur des graphes de connaissances. Une contribution centrale est le développement de systèmes capables de traduire des questions en requêtes formelles, notamment pour des données géospatiales, ainsi que la proposition de benchmarks dédiés à leur évaluation [8, 7]. Ces approches ont été étendues à des systèmes hybrides combinant graphes de connaissances et texte non structuré, permettant d'interroger conjointement des sources telles que Wikidata et Wikipédia [3]. Par ailleurs, d'autres travaux portent sur la complétion de graphes de connaissances à l'aide de modèles de langue [5].

Classification et clustering

L'équipe développe des méthodes de classification et de structuration de données textuelles adaptées aux contextes faiblement supervisés. Nos travaux récents montrent que les stratégies standard de fine-tuning sont sous-optimales dans ce cadre d'apprentissage. Nous proposons des alternatives permettant d'améliorer significativement les performances sans coût d'annotation supplémentaire [4]. En parallèle, nous développons des approches de clustering explicable visant à structurer des données sans supervision [6]. Ces travaux proposent une approche interactive pour explorer de grands corpus textuels à l'aide de modèles de langue. Les documents sont transformés en résumés afin de faciliter leur organisation en groupes thématiques interprétables, tandis que l'utilisateur peut orienter l'analyse selon ses besoins pour une exploration plus ciblée. L'approche vise ainsi à améliorer l'accès, la navigation et la compréhension de collections documentaires complexes, notamment dans le domaine médical.

Applications du TAL

Plusieurs applications spécifiques ont été développées dans le cadre des travaux de l'équipe. Un premier axe s'intéresse à la configuration de réseaux informatiques à l'aide de modèles de langue, avec pour objectif de transformer la gestion d'infrastructure en une interface conversationnelle et décisionnelle. Pour surmonter les limites des LLM classiques, nous avons conçu un modèle fondé sur une approche de type Graph-RAG [13], capable d'intégrer à la fois des sources textuelles, telles que les RFC, les manuels et les historiques de configuration, ainsi que la topologie dynamique du réseau modélisée sous forme de graphe de connaissances.

Un second axe applicatif porte sur l'identification de l'auteur d'un texte, à la fois sur le plan méthodologique [2] et dans des contextes réels. Ces travaux ont notamment donné lieu à une collaboration en cours avec le projet Enccre : de nombreux articles de l'Encyclopédie de Diderot, D'Alembert et Jaucourt ne sont pas signés en raison de la censure de l'époque. Les méthodes développées par les membres de l'équipe ont été mobilisées pour assister les chercheurs du projet dans la désanonymisation de ces articles.

Une dernière application concerne le domaine de la santé, où environ 80% des données sont contenues dans des notes cliniques non structurées. Nos travaux récents [9] proposent une approche visant à transformer ces documents hétérogènes en profils de patients exploitables pour la recherche et l'aide à la décision. Cette méthode s'appuie sur les LLMs pour réaliser une extraction sélective d'entités pivots (symptômes, diagnostics, traitements), adaptée à des cas d'usage prédéfinis. Les données ainsi structurées sont ensuite analysées par clustering afin d'identifier des groupes homogènes de patients, transformant des comptes rendus bruts en informations cliniques exploitables.

Membres permanents

- Marc BERNARD (MCF)
- Catherine COMBES (MCF)
- Rémi EMONET (PR)
- Mathias GÉRY (MCF HDR)
- Antoine GOURRU (MCF)
- Christophe GRAVIER (PR)
- François JACQUENET (PR)
- Baptiste JEUDY (MCF)
- Christine LARGERON (PR)
- Pierre MARET (PR)
- Fabrice MUHLENBACH (MCF HDR)
- Kamal SINGH (MCF HDR)
- Haifei ZHANG (MCF)

Références

[1] Anthony Dubreuil, Antoine Gourru, Christine Largeron, and Amine Trabelsi. Are stereotypes leading llms' zero-shot stance detection ? In *EMNLP*, 2025.

[2] Antoine Gourru, Julien Velcin, Christophe Gravier, and Julien Jacques. Dynamic gaussian embedding of authors. In *WWW*, 2022.

[3] Kunpeng Guo, Clement Defretiere, Dennis Diefenbach, Christophe Gravier, and Antoine Gourru. Qanswer : Towards question answering search over websites. In *WWWW (Companion)*, 2022.

[4] Kunpeng Guo, Dennis Diefenbach, Antoine Gourru, and Christophe Gravier. Fine-tuning strategies for domain specific question answering under low annotation budget constraints. In *IDA*, 2023.

[5] Kunpeng Guo, Dennis Diefenbach, Antoine Gourru, and Christophe Gravier. Wikidata as a seed for web extraction. In *WWW*, 2023.

[6] Carl Hatoum, Catherine Combes, Virginie Fresse, Christophe Gravier, and Mathieu Orzalesi. Clustering de résumés llm guidés par l'utilisateur : vers une approche constructiviste et réaliste unifiée. In *CORIA*, pages 82–95, 2025.

[7] Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Maria-Aggeliki Pollali, Pierre Maret, and Manolis Koubarakis. The question answering system geoqa2 and a new benchmark for its evaluation. *International Journal of Applied Earth Observation and Geoinformation*, 134 :104203, 2024.

[8] Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Mariangela Pollali, Michail Mitsios, Myrto Tsokanaridou, Manolis Koubarakis, and Pierre Maret. Benchmarking geospatial question answering engines using the dataset geoquestions1089. In *ISWC*, pages 266–284. Springer, 2023.

[9] Corentin Laval, Catherine Combes, Rémi Eyraud, and Virginie Fresse. Identification de profils patients à partir de notes cliniques non structurées. *PFIA*, 2025.

[10] Thibaud Leteno, Antoine Gourru, Charlotte Laclau, Remi Emonet, and Christophe Gravier. Fair text classification with wasserstein independence. In *EMNLP*, 2023.

[11] Thibaud Leteno, Michael Perrot, Charlotte Laclau, Antoine Gourru, and Christophe Gravier. Fair text classification via transferable representations. *Journal of Machine Learning Research*, 26(239) :1–47, 2025.

[12] Thibaud Leteno, Irina Proskurina, Antoine Gourru, Julien Velcin, Charlotte Laclau, Guillaume Metzler, and Christophe Gravier. Histoires morales : A french dataset for assessing moral alignment. In *NAACL*, 2025.

[13] Yasmine Ouni, Kamal Singh, Antoine Gourru, and Farouk Mhamdi. Wip : Large language models for network automation : A hierarchical retrieval-augmented approach. In *WoWMoM*, 2026.

[14] Clovis Varangot-Reille, Christophe Bouvard, and Antoine Gourru. Generalising llm routing using past performance retrieval : A few-shot router is sufficient. In *EACL (Student Research Workshop)*, 2026.

[15] Clovis Varangot-Reille, Christophe Bouvard, Antoine Gourru, Mathieu Ciancone, Marion Schaeffer, and François Jacquenet. Doing more with less–implementing routing strategies in large language model-based systems : An extended survey. *JAIR*, 2026.

AfIA
Association française
pour l'Intelligence Artificielle

LaBRI : Laboratoire Bordelais de Recherche en Informatique

LaBRI UMR 5800
CNRS, Université de Bordeaux, Bordeaux INP
https://www.labri.fr

Jean-Luc ROUAS
jean-luc.rouas@labri.fr

Présentation générale

Le LaBRI est une unité de recherche associée au CNRS (UMR 5800), à l'université de Bordeaux et à Bordeaux INP. Depuis 2002, il est partenaire de l'Inria. Ses effectifs se montent à un total d'environ 300 personnes, dont 110 enseignants chercheurs, 40 chercheurs, 20 personnels administratifs et techniques et plus de 100 doctorants, post-doctorants et ingénieurs contractuels. Le LaBRI est structuré en 14 équipes de recherche regroupées en 5 départements, ainsi que deux axes transverses (IA et Santé Numérique).

La recherche en traitement et analyse de la parole se concentre au sein du département "Image et Son", plus précisément dans l'équipe "Traitement et Analyse de Données" (TAD) et s'intègre aux deux axes transverses du laboratoire.

Principaux thèmes de recherche

Les recherches effectuées en traitement de la parole dans l'équipe TAD s'articulent autour de l'analyse, avec une attention particulière portée à l'expressivité véhiculée par les signaux vocaux. Ces derniers révèlent bien plus que les mots prononcés : leur intonation, leur rythme et leur modulation peuvent exprimer, volontairement ou non, des émotions, des états mentaux, ou des attitudes spécifiques. Par exemple, un chanteur peut transmettre un sentiment par son interprétation, tandis qu'un locuteur peut révéler inconsciemment une émotion ou une condition pathologique à travers sa voix. La richesse de ces informations rend l'analyse vocale précieuse, notamment en médecine, où elle peut fournir des indices complémentaires pour évaluer l'état mental ou identifier des pathologies.

Les recherches effectuées par l'équipe ont la particularité de s'étendre sur l'ensemble de la chaîne d'analyse de la parole : depuis les enregistrements jusqu'à l'évaluation finale des systèmes automatiques, en incluant la conception d'expériences perceptives, l'extraction de caractéristiques acoustiques, et l'utilisation de l'apprentissage automatique pour construire des modèles robustes. Cette démarche, fondamentalement interdisciplinaire, a bénéficié de collaborations avec des experts en linguistique, sciences cognitives, mathématiques, apprentissage automatique et sciences médicales.

Les sections suivantes décrivent les compétences de l'équipe via des exemples sur chaque partie de la chaîne d'analyse.

Captation de la voix parlée et chantée

Notre équipe a développé une expertise dans la collecte de données vocales, en ciblant des contextes variés — du chant professionnel aux interactions sociales, en passant par des protocoles cliniques. Ces bases de données, potentiellement multilingues (français, japonais) et multimodales, permettent d'explorer des phénomènes tels que les modes phonatoires, les affects sociaux, ou encore les marqueurs vocaux de pathologies. Cette étape essentielle s'accompagne de défis majeurs, tant sur le plan technique que réglementaire. Notre équipe a su développer une expertise pour surmonter ces obstacles, garantissant à la fois la qualité scientifique des données et leur conformité éthique et légale. Voici quelques exemples marquants de nos travaux :

- Voix chantée : étude des modes phonatoires produits par un chanteur professionnel [7] ;
- Production d'affects sociaux en français et en japonais (une trentaine de locuteurs) [5] ;
- Mise au point d'un jeu collaboratif en réalité virtuelle pour la collecte du rire social (12 locuteurs français et japonais) [19] ;
- Enregistrement au cours d'un protocole clinique au CHU de Bordeaux de sujets suspectés de troubles du sommeil, tâche de lecture (120 participants) [11], parole spontanée (60 sujets) [1] et

de sujets sains en privation de sommeil (lecture, 30 sujets) [10] ;
- Parole empathique (personnel soignants en EHPAD) [17], collecte en conditions simulées et réelles (environ 120 sujets)
- Troubles de l'addiction (collecte en cours), en condition de laboratoire (induction du craving) et conditions réelles (collecte via smartphone)

Ces bases de données servent de fondement à nos recherches en analyse acoustique, modélisation perceptive et apprentissage automatique, avec des applications potentielles en santé, en interaction homme-machine et en sciences sociales.

Experiences de perception

Pour valider la pertinence de nos données et de nos hypothèses, nous menons des expériences de perception auprès de participants humains. Ces études permettent de corréler les caractéristiques acoustiques objectives avec la perception subjective, offrant ainsi une vision holistique des phénomènes étudiés. Parmi nos travaux :

- Combinaison d'indices audio et visuels pour la reconnaissance des affects sociaux mettant en lumière l'importance de la multimodalité dans la communication émotionnelle [18]
- Perception catégorielle du rire : distinction entre rires spontanés et sociaux, avec des implications pour la compréhension des mécanismes sociaux [19]
- Étude de la perception vocale des troubles du sommeil : évaluation de la capacité des auditeurs à détecter la fatigue ou la somnolence à partir de la voix, avec des applications pour le diagnostic précoce [8, 9]

Etude de caractéristiques acoustiques

L'analyse fine des paramètres acoustiques est au cœur de notre démarche pour développer des systèmes de classification automatique. Nous extrayons et modélisons des descripteurs adaptés à chaque phénomène étudié, en combinant méthodes traditionnelles (comme l'analyse spectrale) et approches innovantes (apprentissage profond). Nos contributions incluent :

- Caractérisation des modes phonatoires en voix chantée, avec des applications en ethnomusicologie pour la classification automatique d'extraits vocaux [15]
- Description des affects sociaux via des paramètres vocaux, comparée aux résultats des expériences de perception pour une validation croisée [16]
- Classification des troubles du sommeil à partir de marqueurs acoustiques, ouvrant la voie à des outils d'aide au diagnostic [12]

Apprentissage automatique

Notre équipe est à la pointe de l'apprentissage automatique appliqué à la voix et à la parole, avec des contributions à la fois théoriques et appliquées. Nous développons des modèles end-to-end, optimisons leur déploiement matériel, et explorons des architectures innovantes pour des tâches variées :

- Reconnaissance automatique de la parole en français : développement de modèles performants et contributions à la boîte à outils ESPNET, utilisée mondialement [3, 6]
- Traduction automatique enrichie : intégration d'informations émotionnelles issues de l'analyse de la parole pour améliorer la qualité des traductions [4]
- Co-optimisation matériel/logiciel : adaptation des réseaux de neurones pour des applications embarquées (reconnaissance de parole, traduction), avec un focus sur l'efficacité énergétique et la latence réduite [2, 13, 14]

Membres Impliqués

- Charles BRAZIER (MCF)
- Jean-Luc ROUAS (CR CNRS)
- Georgeta BORDEA (membre associé)
- Takaaki SHOCHI (membre associé)

Références

[1] Colleen Beaumard, Vincent P. Martin, Charles Brazier, Julien Coelho, Jean-Luc Rouas, and Pierre Philip. Medispeech : A French Reading and Spontaneous Speech Corpus for Sleepiness Estimation. In *LREC 2026*, 2026.

[2] Leila Ben Letaifa and Jean-Luc Rouas. Variable Scale Pruning for Transformer Model

Compression in End-to-End Speech Recognition. *Algorithms*, 16(9) :398, September 2023.

[3] Florian Boyer. *Reconnaissance de Parole Pour Le Français et Intégration Dans Un Système de Compréhension Du Langage Parlé*. PhD thesis, Université de Bordeaux, 2021.

[4] Charles Brazier and Jean-Luc Rouas. Conditioning LLMs with Emotion in Neural Machine Translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, Bangkok, Thailand, Thailand, August 2024.

[5] Dominique Fourer, Marine Guerry, Takaaki Shochi, Jean-Luc Rouas, Jean-Julien Aucouturier, and Albert Rilliard. Analyse prosodique des affects sociaux dans l'interaction face à face en japonais. In *XXXèmes Journées d'études Sur La Parole*, Le Mans, France, June 2014.

[6] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, Jing Shi, Shinji Watanabe, Kun Wei, Wangyou Zhang, and Yuekai Zhang. Recent Developments on Espnet Toolkit Boosted By Conformer. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878, June 2021.

[7] Léonidas Ioannidis, Jean-Luc Rouas, and Myriam Desainte-Catherine. Caractérisation et classification automatique des modes phonatoires en voix chantée. In *XXXèmes Journées d'études Sur La Parole*, Le Mans, France, June 2014.

[8] Vincent P. Martin, Aymeric Ferron, Jean-Luc Rouas, and Pierre Philip. "Prediction of Sleepiness Ratings from Voice by Man and Machine" : A Perceptual Experiment Replication Study. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023.

[9] Vincent P Martin, Salin Nathan, Beaumard Colleen, and Jean-Luc Rouas. How does human hearing estimates sleepiness from speech ? In Yiya Chen, Aoju Chen, and Amalia Arvaniti, editors, *Speech Prosody 2024*, pages 160–164, Leiden, Netherlands, July 2024. ISCA.

[10] Vincent P. Martin, Jean-Luc Rouas, Colleen Beaumard, and Pierre Philip. SOMVOICE : A first dataset to study the effects of sleep deprivation on voice characteristics of healthy French speakers. In *LREC 2026*, 2026.

[11] Vincent P. Martin, Jean-Luc Rouas, Jean-Arthur Micoulaud-Franchi, and Pierre Philip. The objective and subjective sleepiness voice corpora. In *12th Edition of Language Resources and Evaluation Conference.*, pages 6525–6533, Marseille, France, May 2020.

[12] Vincent P. Martin, Jean-Luc Rouas, and Pierre Philip. Automatic detection of sleepiness-related symptoms and syndromes using voice and speech biomarkers. *Biomedical Signal Processing and Control*, 91 :105989, May 2024.

[13] Pedro Palacios Almendros and Rafael Medina Morillas and Jean-Luc Rouas and Giovanni Ansaloni and David Atienza Alonso. Systolic Arrays and Structured Pruning Co-design for Efficient Transformers in Edge Systems. In *Great Lakes Symposium on VLSI (GLSVLSI) 2025*, volume 1, pages 1145–1149. Royal Society of Chemistry (RSC), 2025.

[14] Jean-Luc Rouas, Charles Brazier, Leila Ben Letaifa, Rafael Medina, Pedro Palacios, David Atienza, and Giovanni Ansaloni. Structured pruning for efficient systolic array accelerated cascade Speech-to-Text Translation. In *Proc. Interspeech 2025*, pages 893–897, 2025.

[15] Jean-Luc Rouas and Léonidas Ioannidis. Automatic classification of phonation modes in singing voice : Towards singing style characterisation and application to ethnomusicological recordings. In *Interspeech*, volume 2016, pages 150–154, San francisco, United States, September 2016.

[16] Jean-Luc Rouas, Takaaki Shochi, Marine Guerry, and Albert Rilliard. Categorisation of spoken social affects in Japanese : Human vs. machine. In *International Congress of Phonetic Sciences ICPhS 2019*, Melbourne, Australia, August 2019.

[17] Jean-Luc Rouas, Yaru Wu, and Takaaki Shochi. A study on caregivers speech in retirement homes. In Radek Skarnitzl and Jan Volín, editors, *ICPhS 2023 - 20th International Congress of Phonetic Sciences*, pages 2303–2308, Prague, Czech Republic, August 2023. Guarant International.

[18] Takaaki Shochi, Marine Guerry, Albert Rilliard, Erickson Donna, and Jean-Luc Rouas. The combined perception of socio-affective prosody : Cultural differences in pattern matching. *The Journal of the Phonetic Society of Japan*, 24 :84–96, December 2020.

[19] Takaaki Shochi, Marine Guerry, Jean-Luc Rouas, Marie Chaumont, Toyoaki Nishida, and Yoshimasa Ohmoto. Perceptual and acoustic correlates of spontaneous vs. social laughter. In *Proc. 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, Skovde, Sweden, August 2017.

AfIA
Association française
pour l'Intelligence Artificielle

LaTTICe : Langues, Textes, Traitements Informatiques, Cognition

LaTTICe UMR 8094
CNRS, École normale supérieure/PSL et
Université Sorbonne Nouvelle
https://www.lattice.cnrs.fr

Pascal AMSILI
pascal.amsili@ens.fr

Présentation générale

Le LATTICE (Langues, Textes, Traitements informatiques, Cognition) est un laboratoire CNRS associé à l'ENS-PSL et à l'Université Sorbonne Nouvelle, hébergé dans les locaux de l'École normale supérieure. Depuis sa création, le laboratoire a développé une approche résolument pluridisciplinaire, à l'interface des différents domaines évoqués dans son acronyme. Les recherches en traitement automatique des langues (TAL), et plus généralement en Intelligence Artificielle et en Humanités Numériques ont toujours occupé une place importante dans les activités du laboratoire, à la fois en interaction avec les autres thématiques plus proprement linguistiques (description sémantique, linguistique historique, typologie des langues, *etc.*), mais aussi sur des thématiques propres du TAL. Cette situation est reflétée par les quatre thématiques qui structurent l'axe TAL du laboratoire pour le quinquennal actuel :

- Modélisation informatique pour la linguistique
- Données et algorithmes pour le traitement automatique des langues
- Traitement automatique de la diachronie
- Traitement automatique des langues pour les humanités numériques

Principaux thèmes de recherche

Les quelques paragraphes qui suivent illustrent sans souci d'exhaustivité quelques activités récentes menées dans le laboratoire en lien avec le TAL.

Constitution de ressources pour les TAL et les Humanités numériques

Le besoin de ressources annotées de qualité reste important, même pour le français et plus sûrement pour les langues peu dotées. C'est une pratique ancienne dans le laboratoire que la constitution et la mise à disposition de telles ressources. On peut citer le corpus Democrat [4], corpus consacré à la coréférence en français et qui comporte 200 000 expressions référentielles annotées manuellement, formant ainsi le principal corpus d'entraînement pour la résolution de la co-référence.

Outre Democrat, le Lattice a récemment développé d'autres corpus, comme le corpus Cidre (the Corpus for Idiolectal Research) [9]. Cidre est composé d'œuvres de fiction de 11 auteurs français prolifiques du XIXe siècle (4 femmes, 7 hommes ; 22-62 œuvres/auteur ; total de 37 millions de mots). Ce corpus doit permettre d'aborder l'étude de l'évolution de la langue d'un individu (évolution de l'idiolecte ou stylochronométrie, quand c'est l'évolution du style qui est en jeu).

En 2024, le laboratoire a aussi publié le corpus FReND [6], le plus grand corpus annoté pour la négation en français, qui comporte presque 600 000 tokens et environ 9 000 négations annotées en qualité gold.

Le laboratoire développe enfin le versant français du projet Multilingual **BookNLP**, nommé **PROPP** (Pattern Recognition and Ontologies for Prose Processing). BookNLP est à l'origine un projet de l'Université de Berkeley, visant à développer des outils de TAL adaptés à la littérature : les outils sont en effet conçus autour de tâches génériques (analyse morphosyntaxique, syntaxique, etc.), utiles mais peu adaptées au genre romanesque (la notion d'entité nommée est par exemple très insuffisante face à la notion de personnage). Or, l'expérience a montré qu'il était possible d'obtenir des outils efficaces pour l'étude de grands corpus littéraires, au moins pour l'anglais.

Le projet vise à concevoir des outils d'annotation similaires pour le français. Actuellement, une vingtaine d'extraits de romans (du 19e et début 20e, libres de droit, environ 170.000 mots) ont été annotés au Lattice suivant le modèle BookNLP ré-

visé (et en reprenant l'annotation en chaîne de coréférence du projet ANR Democrat, qui a constitué une base solide pour l'annotation). La mise au point des modèles d'annotation, l'évaluation de ces modèles et leur amélioration éventuelle avec des techniques d'apprentissage avancé permettent d'obtenir à ce jour des performances comparables à celles de l'anglais [1].

Approches critiques des modèles de langue

L'émergence des modèles de langue préentraînés rend nécessaire la mise au point de méthodes d'analyse ou de sondage du fonctionnement et des limites des modèles de langue, avec des approches qui peuvent aller d'analyses « comportementales » des modèles vus comme des boîtes noires, à des interventions ciblées dans les représentations internes des modèles, pour identifier les éléments pertinents parmi les milliards de paramètres des modèles, et pour mieux contrôler leur fonctionnement.

On peut citer, dans cette veine, les travaux de thèse de Karim Lasri [5] qui portaient sur la compétence en matière de genre grammatical des modèles de type BERT, ou les travaux de David Kletz [3, 2] qui étudient la façon dont les modèles *transformers* représentent la négation linguistique. Des travaux encore en cours portent sur la mémorisation verbatim dans les grands modèles de langue [11] ou sur les biais de genre dans les LLM [10].

Parsing en dépendance

L'analyse en dépendances est une composante essentielle de nombreuses applications de TAL, dans la mesure où il s'agit de fournir une analyse des relations entre les principaux éléments de la phrase. La plupart des systèmes d'analyse en dépendances sont issus de techniques d'apprentissage supervisées, à partir de grands corpus annotés. Ce type d'analyse est dès lors limité à quelques langues seulement, qui disposent des ressources adéquates. Pour résoudre ce problème, KyungTae Lim, dans sa thèse [7], a examiné plusieurs méthodes d'amorçage, dont l'apprentissage par transfert multilingue, qui a inspiré sa proposition. Son analyseur syntaxique a été évalué sur une soixantaine de langues à travers la participation aux campagnes d'évaluation proposées dans le cadre de la conférence CoNLL. Le système du Lattice a obtenu des résultats très compétitifs lors de campagnes d'évaluation officielles, notamment lors des campagnes CoNLL 2017 et 2018. Ce travail offre donc des perspectives intéressantes pour le traitement automatique des langues peu dotées, un enjeu majeur pour le TAL dans les années à venir [8].

Nos travaux sur le parsing en dépendance, en particulier pour les langues peu dotées, se poursuivent avec un projet récemment financé par Sorbonne Alliance, le projet **TARTOAS** (Traitement automatique des langues régionales : Transcription de l'oral et annotation syntaxique) qui vise à constituer et enrichir des corpus de langues régionales françaises (wallon, picard, occitan, paicî), annotés selon les normes Universal Dependencies. Il inclut la transcription automatique, l'étude des variations orales/écrites, et le développement d'outils TAL pour langues sous-dotées.

Membres permanents

- Pascal AMSILI (PR Sorbonne Nouvelle)
- Mathieu DEHOUCK (CR CNRS)
- Yoann DUPONT (MCF Sorbonne Nouvelle)
- Ioana GALLERON (PR Sorbonne Nouvelle)
- David KLETZ (MCF Sorbonne Nouvelle)
- Frédéric LANDRAGIN (DR CNRS)
- Dominique LEGALLOIS (PR Sorb. Nouvelle)
- Frédrique MÉLANIE-BEQUET (IR CNRS)
- Thierry POIBEAU (DR CNRS)
- Olga SEMINCK (IR CNRS)
- Qi WANG (IE CNRS)

Références

[1] Antoine Bourgois and Thierry Poibeau. The Elephant in the Coreference Room : Resolving Coreference in Full-Length French Fiction Works. In *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 55–69, Suzhou, France, November 2025. Association for Computational Linguistics.

[2] David Kletz, Pascal Amsili, and Marie Candito. The self-contained negation test set.

In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi, editors, *Proceedings of the 6th BlackboxNLP Workshop : Analyzing and Interpreting Neural Networks for NLP*, pages 212–221, Singapore, December 2023. Association for Computational Linguistics.

[3] David Kletz, Pascal Amsili, and Marie Candito. Polarity inversion operators in PLM. In Gemma Boleda and Michael Roth, editors, *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 312–322, Vienna, Austria, July 2025. Association for Computational Linguistics.

[4] Frédéric Landragin. Le corpus Democrat et son exploitation. Présentation. *Langages*, (224) :11–24, December 2021.

[5] Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. Probing for the Usage of Grammatical Number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics 2022*, volume Volume 1 : Long Papers of *ACL (Actes de la 60e conf de l'Association for Computational Linguistics)*, pages 8818–8831, Dublin, Ireland, May 2022. ACL, ACL.

[6] Hafida Le Cloarec Ait Yahya, Olga Seminck, and Pascal Amsili. FReND : A French resource of negation data. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7461–7468, Torino, Italia, May 2024. ELRA and ICCL.

[7] Kyungtae Lim. *Bootstrap methods for multi-task dependency parsing in low-resource conditions*. Theses, Université Paris sciences et lettres, February 2020.

[8] Kyungtae Lim, Niko Partanen, and Thierry Poibeau. Multilingual Dependency Parsing for Low-Resource Languages : Case Studies on North Saami and Komi-Zyrian. In *LREC 2018 Proceedings*, Miyazaki, Japan, May 2018. ELRA.

[9] Olga Seminck, Philippe Gambette, Dominique Legallois, and Thierry Poibeau. The Corpus for Idiolectal Research (CIDRE). *Journal of Open Humanities Data*, 7 :15, 2021.

[10] Laurine Vianne, Yoann Dupont, and Jean Barré. Gender Bias in French Literature. In *Conference on Computational Humanities Research CHR2023*, Paris, France, December 2023. Ariane and Epita and Humanistica.

[11] Xinhao Zhang, Olga Seminck, and Pascal Amsili. Remember to forget : A study on verbatim memorization of literature in large language models. In *Proceedings of the Fifth Conference on Computational Humanities Research (CHR 2024)*, Aahrus, Denmark, 2024. DIGHUMLAB and Center for Humanities Computing, Aahrus University, CEUR Workshop Proceedings.

AfIA
Association française
pour l'Intelligence Artificielle

LIA : Laboratoire Informatique d'Avignon

LIA UPR 4128
Avignon Université
https://lia.univ-avignon.fr

Corinne FREDOUILLE
Coordinatrice du bulletin au LIA
corinne.fredouille@univ-avignon.fr

Fabrice LEFÈVRE
Directeur du LIA (2025-2030)
fabrice.lefevre@univ-avignon.fr

Mots-clés

Langage oral et écrit, interactions vocales, locuteur, réseaux complexes, partenariats industriels, domaine de la santé, défense, éducation.

Le LIA, les TLH et l'IA

Le Laboratoire Informatique d'Avignon (LIA) a été fondé en 1987 par le Professeur Henri MÉLONI, précédemment chercheur du groupe d'intelligence artificielle de Luminy, à Marseille, et l'un des pères fondateurs du langage Prolog. À cette époque, les activités de recherche du laboratoire étaient dédiées exclusivement au traitement de la parole, impliquant des approches à base de règles, de la modélisation logique et symbolique, et des méthodes d'apprentissage automatique (avec, déjà, des réseaux de neurones). Les activités du LIA se sont progressivement étendues, au cours des années, vers le traitement automatique du langage (TAL) au sens large, alliant oral, écrit et traces dans les réseaux sociaux. Elles constituent l'un des thèmes phares du laboratoire. Aujourd'hui, l'apprentissage profond est au cœur d'une grande partie de ses travaux de recherche en langage. Les domaines d'application sont nombreux - santé, défense, sécurité civile, éducation, etc. Les activités du LIA sont largement soutenues par des contrats de collaboration avec des industriels locaux et nationaux, incluant start-up, PME et grands groupes industriels, se traduisant par de nombreuses thèses CIFRE, des contrats d'alternance localisés au laboratoire ou encore le montage de projets tels que la chaire partenariale LIAvignon.

Le laboratoire, qui compte aujourd'hui d'autres thèmes de recherche comme l'étude des systèmes complexes, les réseaux et la recherche opérationnelle, se compose d'un peu plus d'une trentaine de personnels enseignants-chercheurs et chercheurs associés, et d'une quarantaine de doctorants et post-doctorants, répartis en deux équipes Speech and Language Group (SLG) et Complex systems, Operations Research for Networks and Text group (CORNET). Il s'appuie sur deux personnels ingénieurs permanents pour la gestion de son cluster informatique et l'accompagnement à la recherche.

La thématique Langage

La thématique Langage du LIA couvre un large faisceau d'activités, dont quelques exemples de champs d'application sont donnés ci-après. Au travers de cette thématique, le LIA est membre de l'Institut Convergence ILCB et de l'Institut Carnot Cognition.

Traitement automatique de la parole : transcription, traduction, encodage sémantique et paralinguistique de la parole

La thématique de la reconnaissance automatique de la parole existe au LIA depuis sa création. Avec l'essor de l'apprentissage profond et des approches purement neuronales, dites de bout en bout, cette thématique de recherche s'est étendue au LIA à la traduction automatique, la compréhension du langage parlé et bien d'autres tâches qui s'appuient sur l'extraction d'information sémantique dans la parole. Parmi les activités émergentes, nous pouvons citer la pseudonymisation de la parole qui consiste à modifier le contenu linguistique d'enregistrements de parole pour remplacer les informations sensibles par des informations neutres.

Le LIA s'est fortement investi ces dernières années dans l'apprentissage auto-supervisé pour le traitement de la parole. Il est un des acteurs ma-

jeurs du projet collaboratif LeBenchmark qui a permis la création des premiers modèles de fondation appris de manière auto-supervisée sur du français parlé [35]. Actuellement, le LIA participe au projet Pantagruel qui construit des modèles encore plus performants.

Le LIA travaille par ailleurs sur de nombreuses langues et a développé les modèles SENSE qui permettent un encodage sémantique multilingue et bi-modal (parole/texte) de phrases [30]. Ces approches donnent d'excellents résultats, parmi les meilleurs, dans le cadre du traitement des langues peu dotées. En particulier, le LIA travaille sur les dialectes arabes et certaines langues africaines comme le Fongbé ou le Tamasheq.

Dans ce contexte de langues peu dotées, le LIA a su également développer des approches plus traditionnelles, comme par exemple dans ses travaux dédiés au Nawatl (ou Nahuatl), une langue amérindienne vivante, autochtone et langue nationale du Mexique, connue pour son caractère agglutinatif et polysynthétique. Dans ce cadre, le LIA a développé un nouveau corpus [19] et des protocoles d'évaluation sémantique. Il a également introduit deux grammaires formelles basées sur les structures linguistiques nahuatl [20], ainsi que de nouvelles stratégies d'expansion de corpus [21].

Cette expertise autour du traitement automatique de la parole, couplée à celle de la caractérisation du locuteur et de sa voix (développée plus loin), a conduit le LIA à coordonner le projet ANR Le Petit Camion en collaboration étroite avec le SDIS31 et l'Institut de Recherche en Informatique de Toulouse (IRIT). Ce projet est dédié au traitement des appels d'urgence des centres de traitement d'alerte en vue d'optimiser leur prise en charge et d'assister l'opérateur ou opératrice dans sa prise de décision.

Enfin, il est important de souligner la contribution majeure à l'outil *open source* SpeechBrain, très populaire dans la communauté scientifique internationale du traitement de la parole par approches neuronales , dans lequel les membres du LIA sont impliqués depuis sa création.

Traitement automatique du texte et de la parole dans le domaine médical

Parmi les champs d'application du traitement automatique du langage et de la parole au LIA, le domaine médical occupe une place croissante. Les besoins y sont considérables (structuration des dossiers patients, transcription des consultations, aide à la décision clinique, allègement des tâches médico-administratives) mais la nature sensible des données soulève un double défi : leur accès restreint limite les volumes exploitables pour l'entraînement, et les exigences de confidentialité et de protection des données rendent nécessaire le déploiement local des modèles au sein des établissements de santé. Le LIA s'est fortement investi sur ces questions, à la fois sur le versant textuel et sur le versant parole.

Le laboratoire a ainsi contribué à la construction de plusieurs modèles de fondation souverains pour le français médical. DrBERT [24] est le premier modèle de langue de type BERT pré-entraîné pour les domaines biomédical et clinique en français, évalué sur un large ensemble de tâches (reconnaissance d'entités nommées, classification, question-réponse à choix multiples, etc.). Dans la lignée des grands modèles de langue génératifs, BioMistral [25] (fondé sur Mistral 7B) est le premier LLM biomédical français et multilingue en accès libre.

Le laboratoire mène en parallèle des travaux en parole. D'une part, il s'attache à l'adaptation des systèmes de reconnaissance automatique de la parole au domaine médical, marqué par un vocabulaire spécialisé et des conditions bruitées (dictée en consultation, comptes rendus). D'autre part, il étudie la préservation du contenu médical par les systèmes de synthèse vocale ouverts [36], enjeu critique lorsque des termes cliniques doivent être restitués sans altération.

Ces activités s'inscrivent dans plusieurs projets d'envergure nationale. Le projet ANR MALADES, coordonné par Nantes Université (LS2N) et lauréat de l'appel ANR TSIA « Giga-modèles ». Le projet PARTAGES, retenu dans le cadre de France 2030 (Bpifrance) et porté par le Health Data Hub, réunit une trentaine de partenaires (établissements de santé, laboratoires, industriels) autour de la conception d'un grand modèle de langue dédié à la santé et d'outils associés (pseudonymisation, structuration

de données, assistance médico-administrative).

Enfin, ces recherches bénéficient de collaborations industrielles renforcées, notamment avec les sociétés Zenidoc et EDL.

Interaction vocale et agents conversationnels

L'interaction vocale entre l'humain et la machine est un défi de grande importance pour un large nombre de systèmes d'accès à l'information (web, serveurs vocaux, applications mobiles ...) ou en robotique (robots compagnons, *etc.*). Nous allons non seulement vers des systèmes plus performants, mais aussi plus naturels et plus interactifs qui prennent mieux en compte l'utilisateur [26]. Les questions que nous abordons dans cette activité sont par exemple : l'apprentissage en ligne d'agents conversationnels, le dialogue situé ou encore le développement d'interfaces emphatiques qui s'adaptent aux états émotionnels des utilisateurs.

Par exemple, dans le cadre du projet ANR μDialBot, les travaux visaient à augmenter les capacités d'interaction d'un robot en contexte situé par une gestion pro-active de ses interactions dialogiques, basée sur ses perceptions audiovisuelles [17]. En partenariat avec l'entreprise ERM, des robots humanoïdes sont déployés en contexte hospitalier à l'hôpital Broca à Paris (Lusage Lab).

En parallèle de cet objectif principal, les capacités d'interaction vocale du dispositif robotique sont accrues par le recours à une technologie d'agent conversationnel à base de réseaux de neurones de type Transformers. On s'est ainsi attachés à développer des versions françaises de modèles de langage pré-entraînés, spécialisés sur des données de dialogue, en les rendant robustes à des entrées de parole fortement bruitées, ou en adaptant leur expression de personnalité [32], ou dans un contexte multilingue pour des langues peu dotées [33].

Une partie de ces travaux s'attache, de son côté, au développement de solutions permettant le dialogue multi-domaine, en s'appuyant sur une cohorte de sous-systèmes spécialisés par domaine. Une première approche mise en œuvre utilise les capacités agentiques des LLM [38].

En parallèle, des travaux s'intéressent à la modélisation multimodale de l'état du dialogue basée sur les perceptions audiovisuelles du robot, et à la possibilité d'obtenir un embedding social de cet état permettant sa prise en compte par un LLM multimodal. Il s'agit alors d'inférer les actions du robot (linguistiques mais aussi déplacements) à partir d'un contexte riche pour rendre compte de l'environnement partagé par l'humain et le robot. Ces travaux se poursuivent dans le cadre du PEPR ENSEMBLE avec le projet AnandaBot portant sur un nouveau paradigme d'interaction pour la collaboration humain-(humain/robot). L'objectif principal est de développer une chaîne de traitement multimodale pour les interactions verbales triadiques entre un utilisateur, un agent humain et son assistant robotique.

Voix et identité : authentification, détection des fraudes, anonymisation de la voix

Le LIA est un acteur historique et reconnu dans le domaine de la reconnaissance automatique du locuteur. Il développe et maintient Kiwano , un outil *open source* de vérification du locuteur à la pointe de la technologie, qui propose des recettes standardisées, des modèles pré-entraînés et l'intégration des architectures neuronales les plus récentes, avec un fort accent mis sur la reproductibilité des résultats [37] mais aussi sur l'amélioration des performances de ces modèles (voir récemment [10]).

Une attention particulière est portée à l'efficience des systèmes, selon trois axes complémentaires : la conception de modèles compacts par quantification en très faible précision (*low-bit quantization*) [28], la mesure et la réduction de la consommation énergétique et de l'empreinte carbone des architectures neuronales, à l'entraînement comme à l'inférence [27] et des techniques de sélection des données d'apprentissage adaptées aux contextes à ressources contraintes [9].

Les systèmes développés au laboratoire sont régulièrement confrontés aux campagnes d'évaluation internationales, au premier rang desquelles les campagnes NIST SRE (*Speaker Recognition Evaluation*), qui réunissent les meilleurs laboratoires académiques et industriels autour de données et de protocoles d'évaluation communs, permettant ainsi une comparaison rigoureuse de l'état de l'art.

Le LIA est par ailleurs très investi dans la lutte contre les fraudes visant les systèmes d'authentifi-

cation vocale (*anti-spoofing*). Ces systèmes sont en effet vulnérables à la parole produite par synthèse vocale, à la conversion d'identité vocale ou encore au rejeu d'enregistrements. Le laboratoire participe ainsi au projet ANR BRUEL, qui ambitionne de proposer la première méthodologie d'évaluation et de certification des systèmes d'identification par la voix face à ces attaques. Il encadre actuellement plusieurs thèses portant sur la détection de la parole générée par l'IA, et plusieurs publications ont été produites très récemment dans ce domaine, parmi lesquelles on peut citer [16, 15, 14].

Les travaux menés dans le domaine de la reconnaissance du locuteur dépendante du texte constituent une piste prometteuse pour répondre aux défis posés par les *deepfakes* vocaux. Le laboratoire poursuit activement ses recherches dans cette direction, avec plusieurs publications scientifiques [18] et plusieurs brevets déposés récemment.

En complément de ces activités, le laboratoire est également impliqué dans le projet ANR ROBOVOX, consacré à la reconnaissance robuste du locuteur en conditions adverses, ainsi que dans le projet ANR SpeechPrivacy, dédié à la protection de la vie privée et à l'anonymisation de la parole. Il développe par ailleurs des collaborations industrielles soutenues, notamment avec les sociétés Aday et Airbus.

TAL et phonétique clinique

Les troubles de la parole et de la voix sont définis par des difficultés, voire l'incapacité pour une personne de produire des sons articulés et modulés pour former des mots compréhensibles dans un acte de communication, souvent résultant de maladies neurodégénératives (Parkinson ou sclérose en plaque), d'un accident vasculaire cérébral, d'un cancer ORL ...

Impliqué depuis 2004 dans des travaux de recherche pluridisciplinaires appliqués à la caractérisation et à l'évaluation des troubles de la parole et de la voix, les objectifs principaux du LIA sont ici de répondre à une demande récurrente de la part des personnels cliniques d'outils objectifs d'évaluation du niveau de sévérité des altérations de parole/voix ou du niveau d'intelligibilité des patients et patientes impliqués. Le seul outil actuellement à disposition du personnel clinique est l'évaluation perceptive (« à l'oreille »).

Dans les activités les plus récentes du LIA sur ce thème, nous citerons les travaux autour du développement d'approches neuronales interprétables (1) pour évaluer l'intelligibilité ou la sévérité de la parole pathologique [1], ou encore (2) pour comprendre le comportement des personnes expertes dans leur perception et évaluation de la parole pathologique [31]. La restauration de l'intelligibilité de cette parole atypique est également un sujet d'actualité récent, inscrit dans le cadre du projet ANR OLINPIC (2025-2028), en partenariat avec le Laboratoire Parole et Langage (LPL, coordinateur du projet), l'IRIT, le Laboratoire de NeuroPsychoLinguistique de Toulouse (LNPL) et le CHU de Toulouse. Ce projet s'intéresse également à l'atypicité de la parole des apprenants ou apprenantes langue seconde à des fins d'amélioration de leur apprentissage.

TAL et Science des réseaux

Une part de l'activité concernant le TAL menée au LIA se positionne à l'interface avec la science des réseaux. Cela inclut, d'une part, l'extraction de graphes à partir de documents textuels, utilisés notamment comme modèles narratifs, et, d'autre part, l'exploitation de graphes pour traiter différentes tâches concernant le texte.

Nos travaux portant sur la détection d'abus dans des discussions en ligne ont ainsi montré l'intérêt d'exploiter l'évolution de la structure de la conversation, représentée sous forme de graphe, pour détecter ce type d'évènements [34]. L'avantage concerne à la fois la précision des résultats et la robustesse de la méthode. De plus, la combinaison de cette information structurale avec le contenu textuel échangé permet d'améliorer encore la performance de prédiction, même avec de simples approches de type feature engineering [13]. Afin de basculer sur du representation learning, nous avons proposé des méthodes d'apprentissage de représentations pour graphes entiers [12]. En appliquant par ailleurs des outils standard de plongement de texte, nous avons obtenu une meilleure combinaison des deux modalités d'information, qui a amené un autre gain de performance [11].

Une série de travaux récents et actuels se concentre sur le développement d'outils capables d'identifier des personnages ainsi que leurs interactions, dans des œuvres littéraires longues : romans, pièces de théâtre. Le but est notamment d'en tirer des réseaux de personnages, qui peuvent ensuite être exploités pour traiter des problèmes de plus haut niveau [23]. Ceci implique de traiter des tâches de TAL telles que la reconnaissance d'entités nommées ou la résolution d'alias [5, 7]. Le style littéraire des textes considérés et leur longueur importante réduisent significativement la performance des méthodes standard, et rendent nécessaire la conception de méthodes spécifiques [6], et l'élaboration de corpus adaptés à leur apprentissage et à leur évaluation [29]. De plus, le fait que certains de ces textes sont protégés par le droit d'auteur constitue un obstacle juridique au partage des questions juridiques, pour lequel nous avons dû proposer une solution technique [4]. Sur la base des réseaux extraits, nous avons pu attaquer des problèmes tels que la génération automatique de résumés extractifs de séries TV [8], l'alignement d'adaptations différentes de la même histoire [3], la prédiction du niveau de curiosité des lecteurs d'un roman [22], ou encore la détection de complots dans le théâtre romantique [2].

Membres permanents

- Nathalie CAMELIN (SLG)
- Yannick ESTEVE (SLG)
- Corinne FREDOUILLE (SLG)
- Stéphane HUET (SLG)
- Bassam JABAIAN (SLG)
- Pierre JOURLIN (CORNET)
- Rémy KESSLER (SLG)
- Vincent LABATUT (CORNET)
- Fabrice LEFEVRE (SLG)
- Georges LINARES (SLG)
- Driss MATROUF (SLG)
- Salima MDHAFFAR (SLG)
- Mohamed MORCHID (SLG)
- Pascal NOCERA (SLG)
- Mickaël ROUVIER (SLG)
- Juan-Manuel TORRES (CORNET)

Références

[1] S. Abderrazek, C. Fredouille, A. Ghio, M. Lalain, C. Meunier, M. Balaguer, and V. Woisard. Interpretable assessment of speech intelligibility using deep learning : A case study on speech disorders due to head and neck cancers. In *LREC-COLING 2024*, May 2024.

[2] A. Amalvy, N. Diassinous, and V. Labatut. Dynamique des relations dans Lorenzaccio : modélisation par réseaux complexes. In *3èmes Journées d'informatique théâtrale*, Avignon, 2024.

[3] A. Amalvy, M. Janickyj, S. Mannion, P. MacCarron, and V. Labatut. Interconnected kingdoms : Comparing 'a song of ice and fire' crossmedia adaptations using complex networks. *Social Network Analysis and Mining*, 14 :199, 2024.

[4] A. Amalvy, V. Labatut, X. Bost, and H.-H. Huang. Overcoming copyright barriers in corpus distribution through non-reversible hashing. In *64th Annual Meeting of the Association for Computational Linguistics*, 2026.

[5] A. Amalvy, V. Labatut, and R. Dufour. Learning to rank context for named entity recognition using a synthetic dataset. In *Conference on Empirical Methods in Natural Language Processing*, pages 10372–10382, 2023.

[6] A. Amalvy, V. Labatut, and R. Dufour. The role of global and local context in named entity recognition. In *61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 714–722, 2023.

[7] A. Amalvy, V. Labatut, and R. Dufour. The role of natural language processing tasks in automatic literary character network construction. In *31st International Conference on Computational Linguistics*, pages 8462–8473, 2025.

[8] X. Bost, S. Gueye, V. Labatut, M. Larson, G. Linarès, D. Malinas, and R. Roth. Remembering winter was coming : Character-oriented video summaries of tv series. *Multimedia Tools and Applications*, 78(24) :35373–35399, 2019.

[9] P.-M. Bousquet and M. Rouvier. Improving training datasets for resource-constrained speaker recognition neural networks. In *InterSpeech 2023*, 2023.

[10] P.-M. Bousquet and M. Rouvier. Jeffreys divergence-based regularization of neural network output distribution applied to speaker recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[11] N. Cécillon, V. Labatut, and R. Dufour. Conversation-based multimodal abuse detection through text and graph embeddings. *Springer Computing*, 107 :124, 2025.

[12] N. Cécillon, V. Labatut, R. Dufour, and N. Arınık. Whole-graph representation learning for the classification of signed networks. *IEEE Access*, 12 :151303–151316, 2024.

[13] N. Cécillon, V. Labatut, R. Dufour, and G. Linarès. Abusive language detection in online conversations by combining content-and graph-based features. In *International Workshop on Modeling and Mining Socia-Media Driven Complex Networks (Soc2Net)*, volume 2 of *Frontiers in Big Data*, page 8, 2019.

[14] A.-T. Dao, D. Matrouf, M. Rouvier, and E. Nicholas. A comparison of ssl-based feature extractors and back-end classifiers for spoofing detection : A multi-corpus training and cross-linguistic analysis. In *Odyssey*, 2026.

[15] A.-T. Dao, D. Matrouf, M. Rouvier, and E. Nicholas. Speaker-invariant representation learning for spoofing detection via gradient reversal and information bottleneck. In *Odyssey*, 2026.

[16] H. Daumain, D. Matrouf, K. Khelif, and M. Rouvier. From self-supervised speech models to mixture-of-experts for robust anti-spoofing. In *Odyssey*, 2026.

[17] T. Dhaussy, B. Jabaian, and F. Lefèvre. Interaction acceptance modelling and estimation for a proactive engagement in the context of human-robot interactions. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Paris, France, October 2023. IEEE.

[18] N. Griot, D. Matrouf, R. Blouet, J.-F. Bonastre, and A. Manteco. Unified text and speaker verification using ssl model for text-dependent speaker verification. In *Interspeech*, 2025.

[19] J.-J. Guzmán-Landa, J.-M. Torres-Moreno, M.-L. Avendaño-Garrido, M. Figueroa-Saavedra, L. Quintana-Torres, G. Ranger, C.-E. González-Gallardo, E. Linhares-Pontes, P. Velázquez-Morales, and L.-G. Moreno-Jiménez. π-YALLI : un nouveau corpus pour des modèles de langue nahuatl / Yankuik nawatlahtolkorpus pampa tlahtolmachiotl. In *32ème TALN, volume 1*, pages 802–816, Marseille, France, 6 2025. ATALA.

[20] J.-J. Guzmán-Landa, J.-M. Torres-Moreno, G. Ranger, M. Figueroa-Saavedra, L. Quintana-Torres, C.-E. González-Gallardo, L.-G. Moreno-Jiménez, and M.-L. Avendaño-Garrido. Nawatl context-free grammars for Natural Language Processing. In *LREC*, pages 3333–3342, Palma, Spain, May 2026. ELRA.

[21] J.-J. Guzmán-Landa, J.-M. Torres-Moreno, G. Ranger, M. Figueroa-Saavedra, L. Quintana-Torres, E. Linhares-Pontes, L.-G. Moreno-Jiménez, and M.-L. Avendaño-Garrido. Corpora duplication for nlp in low-resource languages : A case study of nahuatl. In ACL, editor, *ACL Workshop Americas 2026*, page 12pp, California, Jul 2026. ACL.

[22] M. D. S. Hopp, V. Labatut, A. Amalvy, R. Dufour, H. Stone, H. Jach, and K. Murayama. Unveiling information gaps : Persistent homology of topic networks for the prediction of reader curiosity. In *63st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 28121–28132, 2025.

[23] V. Labatut and X. Bost. Extraction and analysis of fictional character networks : A survey. *ACM Computing Surveys*, 52(5) :89, 2019.

[24] Y. Labrak, A. Bazoge, R. Dufour, M. Rouvier, E. Morin, B. Daille, and P.-A. Gourraud. Drbert : A robust pre-trained model in french for biomedical and clinical domains. In *Procee-*

dings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), pages 16207–16221, 2023.

[25] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour. Biomistral : A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics : acl 2024*, pages 5848–5864, 2024.

[26] F. Lefèvre. En route to a better integration and evaluation of social capacities in vocal artificial agents. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents - ISIAA 2017*, pages 15–19, New York, USA, 2017. ACM Press.

[27] H. Leguillier, D. Matrouf, G. Lechien, and M. Rouvier. Assessing the energy and carbon emissions of neural speaker verification model in training and inference. In *Odyssey*, 2026.

[28] H. Leguillier, D. Matrouf, G. Lechien, and M. Rouvier. On low-bit quantization errors in speaker verification : Diagnostic and mitigation. In *Odyssey*, 2026.

[29] P. Maurel, A. Amalvy, V. Labatut, and M. Alrahabi. De l'annotation intégrale à l'analyse des réseaux de personnages : un modèle pour la REN dans les textes littéraires en français. *Traitement Automatique des Langues (TAL)*, 66(1) :37–62, 2025.

[30] S. Mdhaffar, H. Elleuch, C. Chellaf, H. Nguyen, and Y. Estève. Sense models : an open source solution for multilingual and multimodal semantic-based tasks. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8, 2025.

[31] T. Nguyen, C. Fredouille, A. Ghio, M. Lalain, and V. Woisard. What does a pathological speech assessment model know about acoustic features ? a case study on oral and oropharyngeal cancer patients. In *Proceedings of Interspeech'26, Sydney, Australia*, 2026.

[32] A. Njifenjou, V. Sucal, B. Jabaian, and F. Lefèvre. Enabling trait-based personality simulation in conversational LLM agents : Case study of customer assistance in French. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 299–308, Bilbao, Spain, May 2025. Association for Computational Linguistics.

[33] A. Njifenjou, V. Sucal, B. Jabaian, and F. Lefèvre. Open-source large language models as multilingual crowdworkers : Synthesizing open-domain dialogues in several languages with no examples in targets and no machine translation. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 697–749, Avignon, France, August 2025. Association for Computational Linguistics.

[34] E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. Conversational networks for automatic online moderation. *IEEE Transactions on Computational Social Systems*, 6(1) :38–55, 2019.

[35] T. Parcollet, H. Nguyen, S. Evain, M. Z. Boito, A. Pupier, S. Mdhaffar, H. Le, S. Alisamir, N. Tomashenko, M. Dinarelli, et al. Lebenchmark 2.0 : A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, 86 :101622, 2024.

[36] R. Rodriguez, S. Huet, B. Favre, and M. Rouvier. Préservation du contenu médical par des systèmes de synthèse vocale ouverts. In *Atelier TAL et Sant {\'e} 2026*.

[37] M. Rouvier and P.-M. Bousquet. Kiwano : A cutting-edge open-source toolkit for speaker verification. In *Odyssey*, 2026.

[38] V. Sucal, M. Jullien, A. Njifenjou, and F. Lefèvre. Llm-based agentic workflow on verbal and non-verbal audiovisual perceptions and actions for proactive situated human-robot interactions. In *Proceedings of the 13th International Conference on Human-Agent Interaction*, HAI '25, page 177–187, New York, NY, USA, 2026. Association for Computing Machinery.

AfIA
Association française
pour l'Intelligence Artificielle

LIASD/GLAÇON : Groupe, Langage, Apprentissage, Corpus, évaluatiON

LIASD - UR 4383
Université Paris 8
https://glacon.up8.site/

Anna PAPPA
apappa@univ-paris8.fr

Présentation générale

GLAÇON [2] est un groupe de recherche spécialisé en traitement du langage au sein de l'équipe PASTIS [3], rattachée au Laboratoire d'Intelligence Artificielle et Sémantique des Données (LIASD).

Nos recherches se situent à l'interface du traitement automatique des langues, de l'apprentissage automatique et de l'intelligence artificielle, avec une attention particulière portée aux langues minoritaires et aux données langagières issues de contextes réels. Le groupe développe une approche centrée sur le corpus, considérant les données textuelles non seulement comme support d'apprentissage pour les modèles, mais aussi comme objet d'étude à part entière, en lien avec leurs conditions de production, leur structuration et leurs usages.

Les recherches menées au sein de GLAÇON couvrent plusieurs axes complémentaires :

- l'étude et le traitement des langues minoritaires ou peu dotées, avec des enjeux spécifiques liés à la rareté des ressources et à l'adaptation des méthodes ;
- la constitution de corpus multilingues et/ou spécialisés, incluant des problématiques de collecte web, de normalisation et de qualité des données ;
- les méthodes d'annotation, en particulier dans des contextes d'incertitude ou de complexité interprétative, ainsi que l'étude de leur impact sur les modèles ;
- le développement d'approches hybrides pour l'analyse du langage, combinant méthodes statistiques et apprentissage, en particulier avec des modèles de langage pré-entraînés (architectures Transformer, encodeurs ou génératifs), dans la continuité de travaux antérieurs en apprentissage profond appliqué au TAL [9] ;
- l'étude de l'intégration de l'IA générative et des outils TAL dans l'enseignement, avec une attention aux enjeux pédagogiques et éthiques.

Ces travaux s'inscrivent dans une démarche interdisciplinaire, articulant l'informatique aux sciences humaines et sociales. Les collaborations du groupe couvrent la littérature et la stylistique (analyse et modélisation des styles d'écriture), les arts et les humanités numériques (étude des représentations dans les corpus culturels, tels que le cinéma ou le théâtre), ainsi que les sciences économiques et sociales, à travers l'analyse de données textuelles issues de contextes professionnels.

Principaux thèmes de recherche

Les travaux du groupe s'organisent autour d'un axe principal : considérer le langage comme une donnée complexe, à recueillir, annoter, interpréter et mobiliser dans différents contextes réels. Notre champ de recherche couvre l'ensemble du traitement des données langagières, de la collecte à la production de ressources exploitables, en passant par la conception, l'annotation et le traitement de corpus.

Construction et qualité des corpus

Un premier axe concerne la constitution de corpus textuels à partir de sources hétérogènes (web, archives, plateformes numériques), souvent multilingues et bruitées. La qualité, la traçabilité et la contextualisation des données constituent un enjeu central du TAL contemporain, conditionnant directement les performances et la robustesse des modèles. L'ensemble du traitement des données langagières est couvert : de l'acquisition automati-

2. Groupe, Langage, Apprentissage, Corpus, évaluatiON.
3. Programming, Artificial intelligence, Security, Texts, Images, Simulations.

sée (*crawling, scraping*), nettoyage, normalisation et documentation finale. Les corpus thématiques développés intègrent des dimensions sémantiques et discursives, en lien avec les contextes d'usage [17, 2]. Ces travaux ont conduit à la production de ressources multilingues originales dans des contextes interdisciplinaires, comme le corpus MALANTIN dédié à l'étude de l'innovation non technologique, diffusé en accès ouvert [6].

Ils s'accompagnent de la mise à disposition de ressources et d'outils pour la communauté, tels que des corpus annotés et des outils de caractérisation linguistique, dont le corpus NEM.fr et l'adaptation au français de la librairie `pseudobibeR`, qui permettent une analyse fine des variations de genres textuels.

Langues minorisées et variation linguistique

Un second axe porte sur le traitement des langues minorisées de France (le corse [15] et le breton [14]), présentant des variantes dialectales, graphiques ou liées à leurs usages numériques. Ces contextes posent des défis spécifiques liés à la rareté des ressources, à l'absence de normalisation et à la diversité des pratiques langagières. Les travaux développés explorent des approches alternatives de production de ressources, en s'appuyant sur le crowdsourcing, les jeux sérieux et des dispositifs participatifs [13, 16, 12]. Ces méthodes permettent de collecter et d'annoter des données en mobilisant les locuteurs eux-mêmes, tout en intégrant des dimensions sociolinguistiques et culturelles. En parallèle, des travaux portent sur l'évaluation des systèmes de TAL, en analysant l'impact de la variation linguistique (domaines, genres, registres) sur les performances des modèles [11, 10]. Ces analyses mettent en évidence les limites des approches actuelles et soulignent l'importance de la robustesse des modèles face à la variation des genres textuels, en particulier pour des tâches comme la reconnaissance d'entités nommées [20].

Modélisation, interprétation, usages du langage

Le troisième axe porte sur la modélisation des données langagières et leur interprétation, en mobilisant des approches d'apprentissage automatique pour analyser des phénomènes sémantiques complexes (aspects, concepts émergents, représentations sociales). Ces travaux s'appuient sur des stratégies d'annotation avancées (apprentissage actif, transfert, annotation sous incertitude), intégrant la possibilité de multiples interprétations et permettant de traiter des données partiellement annotées ou ambiguës [3, 4, 18]. Ils incluent également des recherches sur la structuration et l'émergence des nouveaux concepts dans les corpus, à travers la construction de lexiques [7]. Dans cet axe s'inscrivent également des travaux sur les représentations latentes des grands modèles de langage et sur les propriétés émergentes qu'ils manifestent dans leurs productions telles que des profils comportementaux spécifiques à chaque modèle, relativement stables et différenciés [21]. Enfin, cet axe intègre l'étude des usages de l'IA générative en contexte éducatif [19, 8].

Ces recherches se prolongent dans le développement de systèmes génératifs fondés sur des corpus spécialisés pour modéliser et reproduire des styles linguistiques. Appliqués à des corpus théâtraux, ils capturent et transfèrent de régularités stylistiques en contexte multilingue [5, 1].

Membres permanents

- Anna Pappa (PR)
- Revekka Kyriakoglou (MCF)
- Alice Millour (MCF)

Références

[1] R. Baz, K. Malyseva, A. Pappa, and T. Cazenave. BrAIcht : A theatrical agent that speaks like bertolt brecht's characters. In *Proceedings of the Generative Art Conference*, pages 289–296, Rome, Italy, 2023.

[2] M. Boudabous and A. Pappa. WebT-IDC : A web tool for intelligent dataset creation, a use case for forums and blogs. *Procedia Computer Science*, 192 :1051–1060, 2021.

[3] M. Boudabous and A. Pappa. Apprentissage actif pour l'extraction des aspects explicites : application à des avis non annotés en français. *(LIFT)TAL*, pages 20–28, 2022.

[4] M. Boudabous and A. Pappa. Explicit aspect annotation via transfer and active learning. *Procedia Computer Science*, 225 :1124–1133, 2023.

[5] T. Cazenave, G. Grosjean, B. Rozière, and A. Pappa. MoliAIre, a theatrical agent which speaks like molière's characters. In *Generative Art Conference*, pages 286–294, 2022.

[6] R. Kyriakoglou and A. Pappa. MALANTIN : Multilingual corpus for non-technological innovation. https://doi.org/10.82270/malantin/v1, 2025. ORTOLANG (Open Resources and Tools for Language).

[7] R. Kyriakoglou, A. Pappa, J. He, and A. et al. Schoen. Développement automatique de lexiques pour les concepts émergents : une exploration méthodologique. In *(JADT)*, 2024.

[8] R. Kyriakoglou, A. Pappa, V. Psyché, S. Lallé, Guilherme Medeiros Machado, N. El Mawas, S. Proust-Androwkha, L. Chartofylaka, and A. Boubaker. From ethical discourse to empirical evidence : How ethical concerns are operationalized in studies on generative ai in higher education. In *International Conference on Artificial Intelligence in Education*, pages 210–224. Springer, 2026.

[9] L. Medrouk and A. Pappa. Do deep networks really need complex modules for multilingual sentiment polarity detection and domain classification ? In *(IJCNN)*. IEEE, 2018.

[10] A. Millour, Y. Dupont, K. Fort, and L. Duignan. Unveiling strengths and weaknesses of nlp systems based on a rich evaluation corpus : the case of ner in french. In *LREC-COLING*, 2024.

[11] A. Millour, Y. Dupont, A. Jouglar, and K. Fort. Fenec : un corpus équilibré pour l'évaluation des entités nommées en français. In *JEP/TALN/RECITAL*. ATALA, 2022.

[12] A. Millour and K. Fort. Krik : First steps into crowdsourcing pos tags for kréyòl gwadloupéyen. In *Workshop on Computational Approaches to Linguistic Code-Switching and Under-Resourced Languages (CCURL)*, 2018.

[13] A. Millour and K. Fort. À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. *TAL*, 59(3) :41–65, 2018.

[14] A. Millour, L. Grobol, W. Zemouri, Y. Drapier, and M. Jouitteau. Prerequisites for advancing automatic speech recognition in breton. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, volume 11, pages 1464–1473, 2026.

[15] A. Millour, L. Kevers, L. Brasile, and A. Ghia. Agettivu, aggitivu o aghjettivu ? POS Tagging Corsican Dialects. In *LREC-COLING 2024*, Turin,, 2024.

[16] L. Nicolas, V. Lyding, C. Borg, C. Forascu, K. Fort, K. Zdravkova, I. Kosem, J. Čibej, Špela Arhar H., and A. et al. Millour. Creating expert knowledge by relying on language learners : a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *(LREC)*, 2020.

[17] A. Pappa. Corpus textuels : enjeux et défis. *1024 Bulletin de la Société Informatique de France*, (25), 2025.

[18] A. Pappa, P. Pomeret-Coquot, F. Faux, and R. Kyriakoglou. Annotation linguistique sous incertitude structurelle : vers un modèle ensembliste basé sur les fonctions de croyance. In *29e Conférence Nationale en Intelligence Artificielle (CNIA 2026)*, Arras, France, June 2026. Plate-Forme Intelligence Artificielle, Association Française pour l'Intelligence Artificielle.

[19] S. Proust-Androwkha, V. Psyché, and R. et al. Kyriakoglou. Beyond programming assistance towards the integration of generative ai into computer science education : A scoping review. In *Artificial Intelligence in Education (AIED)*, 2025.

[20] Z. Xu, A. Millour, C-E. Gonzalez-Gallardo, and J-Y. Antoine. Adaptabilité des grands modèles de langage aux genres linguistiques attestés (agla-gla). In *(JIRC)*, Blois, France, 2025.

[21] C. Zacharopoulos and R. Kyriakoglou. Decoding emergent big five traits in large language models : Temperature-dependent expression and architectural clustering. In *IJCNLP*, 2025.

AfIA
Association française
pour l'Intelligence Artificielle

LIG/GETALP : Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole

LIG UMR 5217 / GETALP
CNRS et Université Grenoble Alpes
https://lig-getalp.imag.fr

Benjamin LECOUTEUX
benjamin.lecouteux@univ-grenoble-alpes.fr

Didier SCHWAB
Responsable d'équipe
didier.schwab@univ-grenoble-alpes.fr

Présentation de GETALP

Créée en 2007 au sein du Laboratoire d'informatique de Grenoble, l'équipe GETALP réunit informaticiens, linguistes, phonéticiens, traductologues et spécialistes du signal autour d'un même objet : la communication humaine, écrite ou orale, dans sa diversité linguistique, sociale et situationnelle. L'essor des modèles préentraînés et de l'IA générative a changé l'échelle des méthodes, sans modifier son principe directeur : partir des usages et des données, construire et évaluer des modèles, puis revenir aux utilisateurs et aux phénomènes observés.

Cette approche prend en compte les conditions réelles et vise ainsi une forte validité écologique : parole spontanée, bruit, données rares, locuteurs atypiques et domaines spécialisés ; agnostique, parce qu'elle articule modèles neuronaux, expertise et ressources symboliques ; et centrée sur l'humain, en plaçant l'explicabilité, l'accessibilité et l'interaction au cœur de l'évaluation. Elle donne ainsi sa cohérence à des recherches qui vont du benchmark de la parole française à la documentation de langues créoles, de la traduction vers des pictogrammes à l'analyse des affects et à l'interopérabilité des dictionnaires.

Modèles et référentiels ouverts pour le français

GETALP développe des modèles de référence, mais aussi des cadres permettant d'analyser ce qu'ils apprennent. LeBenchmark évalue les représentations auto-supervisées de la parole française sur des tâches de reconnaissance, de compréhension, de traduction et d'analyse des émotions [5, 14]. Sa seconde version renforce la standardisation, la reproductibilité et l'étendue des comparaisons [32, 6]. Des analyses syntaxiques examinent également les connaissances linguistiques acquises, dans le cadre du projet SYNPAX, au-delà des seuls scores applicatifs [4].

Cette même exigence vaut pour les langues de spécialité. JARGON réunit des modèles et des tâches couvrant plusieurs domaines professionnels et scientifiques, afin de mesurer les écarts de performance entre le français général et plusieurs domaines professionnels ou scientifiques [34]. Pantagruel propose, quant à lui, des encodeurs auto-supervisés unifiés pour le texte et la parole en français [16]. Ces infrastructures fournissent à la communauté des points de comparaison fiables et aux projets appliqués des briques documentées, ouvertes et réutilisables.

Parole spontanée et langues peu dotées

Les disfluences, les variations acoustiques, les chevauchements et les erreurs de transcription font de la parole spontanée un terrain d'étude particulièrement exigeant. GETALP en propose une typologie à la fois linguistique et opérationnelle [7], puis explore l'analyse syntaxique directement à partir du signal au moyen de modèles de bout en bout [33].

Les langues peu dotées ne sont pas abordées comme des cas marginaux, mais comme des terrains qui révèlent les limites des méthodes. Pour les créoles, GETALP combine la reconnaissance automatique, la recherche d'exemples parlés et les outils de documentation [25, 17, 18]. L'exploitation d'enregistrements de terrain numérisés suppose de caractériser leurs conditions de collecte et leur adéquation aux modèles actuels [9, 13, 12]. Dans ce cadre, l'équipe a notamment proposé des modèles auto-supervisés adaptés au créole haïtien [11, 10].

Cette démarche associe étroitement la modélisation au travail mené avec les linguistes : choix des unités, annotation, constitution des corpus, mesure de la qualité et retour vers les communautés productrices des données. Elle montre qu'un modèle multilingue massif ne suffit pas à garantir de bonnes performances : la proximité des données, leur qualité documentaire et la définition de la tâche restent déterminantes.

Traduire entre langues, modalités et capacités

La traduction automatique reste un axe central de GETALP, désormais élargi à la parole et aux modalités non textuelles. Les recherches portent sur le préentraînement pour la traduction de la parole [15], ainsi que sur la prise en compte du contexte par les modèles neuronaux [3] et sur les moyens de rendre leurs décisions inspectables. Le projet MAKE-NMTViz associe ainsi chercheurs et traducteurs pour évaluer l'exactitude, l'intelligibilité et l'utilité des visualisations, sans confondre la visibilité d'un signal interne avec la qualité d'une explication [8].

La traduction de la parole vers des pictogrammes constitue un chantier singulier, développé dans le cadre du projet PROPICTO [27]. GETALP a proposé un corpus multimodal alignant parole, transcription et pictogrammes [21], ainsi que des plateformes dédiées à la création de ces données [20]. Voice2Picto a d'abord concrétisé cette orientation sous la forme d'un système complet [24]. Des approches en cascade et de bout en bout ont ensuite été proposées et comparées [22, 23]. Produire une suite de pictogrammes ne revient en effet ni à résumer une phrase ni à remplacer mécaniquement les mots : il faut sélectionner l'information utile, gérer les implicites et respecter les conventions pictographiques.

La simplification complète cette chaîne. GETALP étudie les stratégies à l'œuvre dans la parole spontanée [31], produit des données de qualité avec l'aide de grands modèles de langue [30] et évalue automatiquement la complexité des phrases [29].

Comprendre et expliquer les modèles

L'explicabilité est abordée à plusieurs niveaux. À l'échelle des usages, les visualisations de traduction sont confrontées aux besoins des utilisateurs [8]. À l'échelle des mécanismes internes, une étude montre que plusieurs circuits, algorithmes ou sous-espaces peuvent rendre compte d'un même comportement : l'explication n'est donc pas nécessairement unique, même dans des cas contrôlés [26]. Ce constat conduit à préciser les propriétés attendues d'une explication, qu'il s'agisse de son pouvoir prédictif, de la possibilité d'intervenir sur le modèle, de sa stabilité ou de son intelligibilité.

Ressources lexicales, web sémantique et connaissances

Le développement des grands modèles ne fait pas disparaître le besoin de ressources linguistiques structurées. GETALP maintient DBnary, qui transforme les informations lexicales de Wiktionary en données liées multilingues. La mise en correspondance des sens entre les langues en améliore la navigation et facilite sa réutilisation [35].

L'équipe contribue également au web sémantique. OntoLex-FrAC relie lexiques et corpus afin de représenter les attestations, les fréquences et les collocations [2]. L'articulation entre graphes de connaissances et modèles génératifs ouvre une voie complémentaire : KGAST produit des textes synthétiques annotés à partir de graphes [37], tandis que la génération guidée par ontologie soutient l'extraction d'information spécialisée dans les domaines disposant de peu de données [36].

Interaction, affects et santé

Les travaux de GETALP sur les affects ne se limitent pas aux émotions discrètes : ils modélisent aussi des dimensions continues, comme l'activation et la valence, dont la manifestation dépend du locuteur, du corpus et de la situation. L'apprentissage multi-corpus et les représentations auto-supervisées visent à généraliser entre différentes conditions d'enregistrement [1]. L'émotion est également envisagée comme un phénomène dynamique, grâce à un apprentissage multitâche qui prédit conjointement l'état affectif et son évolution

temporelle [38]. À l'interface de la parole et de l'interaction, l'équipe évalue par ailleurs l'efficacité des modèles de langue français pour le résumé automatique de dialogues [39].

Dans le domaine de la santé, MedDialog-FR fournit un corpus de dialogues médicaux en français destiné à la classification et à la génération de réponses sur la santé intime des femmes [19]. Les travaux sur la parole spontanée médicale complètent cette orientation en traitant les difficultés de reconnaissance propres au domaine [28]. Plusieurs de ces travaux accordent une attention particulière à la validation humaine, considérée comme une composante de la démarche plutôt que comme une vérification ajoutée après l'entraînement.

Une recherche ouverte et collective

GETALP diffuse ses modèles, corpus, codes et ressources sur des plateformes ouvertes, lorsque les contraintes juridiques, éthiques et contractuelles le permettent. Cette politique ne se limite pas à accompagner les publications : elle structure la conception des benchmarks, la documentation des jeux de données et la reproductibilité des évaluations. LeBenchmark, JARGON, les corpus créoles, les ressources pictographiques et DBnary illustrent ainsi la volonté de transformer les résultats scientifiques en objets réutilisables par la communauté.

Conclusion

GETALP aborde les modèles préentraînés et génératifs tout en préservant sa singularité : le multilinguisme, la parole réelle, les langues peu dotées, la structuration des connaissances et l'inclusion des utilisateurs. Qu'il s'agisse de LeBenchmark, de JARGON, des travaux sur la parole créole et la syntaxe parlée, de la traduction vers les pictogrammes, de MAKE-NMTViz ou d'OntoLex-FrAC, une même ambition se dégage : transformer les avancées méthodologiques en ressources, en protocoles d'évaluation et en dispositifs partageables. Cette capacité à relier modèles, données et usages place l'équipe dans une position originale, attentive aux voix, aux langues et aux situations que le changement d'échelle tend à invisibiliser.

Membres permanents

- Véronique AUBERGÉ (CR)
- Valérie BELLYNCK (MCF)
- Hervé BLANCHON (MCF, HDR)
- Francis BRUNET-MANQUAT (MCF)
- Maxime CALKA (MCF)
- Maximin COAVOUX (CR)
- Marco DINARELLI (CR)
- Emmanuelle ESPERANÇA-RODIER (MCF)
- Benjamin LECOUTEUX (Pr, HDR)
- Claire LEMAIRE (associée, MCF)
- Chuyuan LI (CR)
- Maxime PEYRARD (Professeur Junior CNRS)
- François PORTET (Pr, HDR)
- Fabien RINGEVAL (MCF)
- Solange ROSSATO (MCF)
- Didier SCHWAB (Pr, HDR)
- Gilles SÉRASSET (MCF)

Références

[1] Sina Alisamir, Fabien Ringeval, and François Portet. Multi-corpus Affect Recognition with Emotion Embeddings and Self-Supervised Representations of Speech. In *10th International Conference on Affective Computing and Intelligent Interaction (ACII 2022)*, Nara, 2022. IEEE. hal-03930191.

[2] Christian Chiarcos, Ranka Stanković, Maxim Ionov, and Gilles Sérasset. Bridging Computational Lexicography and Corpus Linguistics : A Query Extension for OntoLex-FrAC. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2504–2514, Torino, 2024. hal-04535067.

[3] Marco Dinarelli, Dimitra Niaouri, Fabien Lopez, Gabriela Gonzalez-Saez, Mariam Nacklé, Emmanuelle Esperança-Rodier, Caroline Rossi, Didier Schwab, and Nicolas Ballier. Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention. *Revue TAL : traitement automatique des langues*, 64(3), 2024. hal-04581509.

[4] Zdravko Dugonjić, Adrien Pupier, Benjamin Lecouteux, and Maximin Coavoux. What has

LeBenchmark Learnt about French Syntax ? In *LREC-COLING 2024*, pages 17493–17499, Torino, 2024. ELRA and ICCL. hal-04592628.

[5] Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *INTERSPEECH 2021 : Conference of the International Speech Communication Association*, Brno, 2021. hal-03317730.

[6] Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. Modèles neuronaux pré-appris par auto-supervision sur des enregistrements de parole en français. In *JEP 2022*, île de Noirmoutier, 2022. hal-03707064.

[7] Solène Evain, Solange Rossato, François Portet, and Benjamin Lecouteux. Typologie de la parole spontanée à des fins d'analyse linguistique et de développement de systèmes de reconnaissance automatique de la parole. In *XXXIVe Journées d'Études sur la Parole – JEP 2022*, île de Noirmoutier, 2022. hal-03713384.

[8] Gabriela Gonzalez-Saez, Fabien Lopez, Mariam Nakhle, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Caroline Rossi, Didier Schwab, Jun Yang, James Robert Turner, and Nicolas Ballier. The MAKE-NMTViz Project : Meaningful, Accurate and Knowledge-limited Explanations of NMT Systems for Translators. In *EAMT : European Association for Machine Translation*, Sheffield, 2024. hal-04638945.

[9] William N. Havard, Renauld Govain, Daphne Gonçalves Teixeira, Benjamin Lecouteux, and Emmanuel Schang. Technologies de la parole et données de terrain : le cas du créole haïtien. In *35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)*, volume 1 : articles longs et prises de position, pages 686–694, Toulouse, 2024. ATALA & AFPC. hal-04623051.

[10] William N. Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. Modèles auto-supervisés de traitement de la parole pour le Créole Haitien. In *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)*, pages 542–554, Marseille, 2025. ATALA & ARIA. hal-05330617.

[11] William N. Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. Self-Supervised Models of Speech Processing for Haitian Creole. In *Interspeech 2025*, pages 4018–4022, Rotterdam, 2025. hal-05074823.

[12] William N. Havard, Shrita Hassamal, Muhsina Alleesaib, Guilhem Florigny, Guillaume Fon Sing, Anne Abeillé, Benjamin Lecouteux, and Emmanuel Schang. Exploration de la séparation en langues dans les modèles de traitement de la parole auto-supervisés multilingues préentraînés avec des données écologiques. In *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)*, pages 390–403, Marseille, 2025. ATALA & ARIA. hal-05330610.

[13] William N. Havard, Emmanuel Schang, and Benjamin Lecouteux. Mobilising the Archive :

Training Modern Speech Technology Models with Digitalised Fieldwork Recordings. In *Recent Advances in Language Documentation and Archiving (LD&A'24)*, Berlin, 2024. hal-05059842.

[14] Hang Le, Sina Alisamir, Marco Dinarelli, Fabien Ringeval, Solène Evain, Ha Nguyen, Marcely Zanon Boito, Salima Mdhaffar, Ziyi Tong, Natalia Tomashenko, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Didier Schwab, and Laurent Besacier. LeBenchmark, un référentiel d'évaluation pour le français oral. In *34e Journées d'étude sur la parole JEP 2022*, île de Noirmoutier, 2022. hal-03767742.

[15] Phuong-Hang Le, Hongyu Gong, Changhan Wang, Juan Pino, Benjamin Lecouteux, and Didier Schwab. Pre-training for Speech Translation : CTC Meets Optimal Transport. In *International Conference on Machine Learning (ICML), 2023*, Honolulu, Hawaii, 2023. hal-04117237.

[16] Phuong-Hang Le, Valentin Pelloin, Arnault Chatelain, Maryem Bouziane, Mohammed Ghennai, Qianwen Guan, Kirill Milintsevich, Salima Mdhaffar, Aidan Mannion, Nils Defauw, Shuyue Gu, Alexandre Audibert, Marco Dinarelli, Yannick Estève, Lorraine Goeuriot, Steffen Lalande, Nicolas Hervé, Maximin Coavoux, François Portet, Étienne Ollion, Marie Candito, Maxime Peyrard, Solange Rossato, Benjamin Lecouteux, Aurélie Nardy, Gilles Sérasset, Vincent Segonne, Solène Evain, Diandra Fabre, and Didier Schwab. Pantagruel : Unified Self-Supervised Encoders for French Text and Speech. In *The Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10168–10191, Palma, 2026. hal-05627744.

[17] Eric Le Ferrand, Fabiola Henri, Benjamin Lecouteux, and Emmanuel Schang. Application of Speech Processes for the Documentation of Kréyòl Gwadloupéyen. In *Field Matters 2023 : The Second Workshop on NLP Applications to Field Linguistics*, Dubrovnik, 2023. hal-04129399.

[18] Eric Le Ferrand, Claudel Pierre-Louis, Ruoran Dong, Benjamin Lecouteux, Daphné Gonçalves-Teixeira, William N. Havard, and Emmanuel Schang. Outiller la documentation des langues créoles. In *LIFT 2023 : journées scientifiques du GdR Linguistique Informatique, Formelle et de Terrain*, Vandoeuvre-Lès-Nancy, 2023. hal-04302623.

[19] Xingyu Liu, Vincent Segonne, Aidan Mannion, Didier Schwab, Lorraine Goeuriot, and François Portet. MedDialog-FR : a French Version of the MedDialog Corpus for Multi-label Classification and Response Generation related to Women's Intimate Health. In *First Workshop on Patient-Oriented Language Processing @ LREC-COLING, 2024*, Turin, 2024. hal-04534338.

[20] Cécile Macaire, Jordan Arrigo, Chloé Dion, Claire Lemaire, Emmanuelle Esperança-Rodier, Benjamin Lecouteux, and Didier Schwab. Plateformes pour la création de données en pictogrammes. In *18e Conférence en Recherche d'Information et Applications, 16e Rencontres Jeunes Chercheurs en RI, 30e Conférence sur le Traitement Automatique des Langues Naturelles, 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 6–9, Paris, 2023. ATALA. hal-04208581.

[21] Cécile Macaire, Chloé Dion, Jordan Arrigo, Claire Lemaire, Emmanuelle Esperança-Rodier, Benjamin Lecouteux, and Didier Schwab. A Multimodal French Corpus of Aligned Speech, Text, and Pictogram Sequences for Speech-to-Pictogram Machine Translation. In *LREC-COLING 2024*, Turin, 2024. hal-04534234.

[22] Cécile Macaire, Chloé Dion, Didier Schwab, Benjamin Lecouteux, and Emmanuelle Esperança-Rodier. Approches cascade et de bout-en-bout pour la traduction automatique de la parole en pictogrammes. In *35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automa-*

tique des Langues (RECITAL 2024), volume 1 : articles longs et prises de position, pages 22–35, Toulouse, 2024. ATALA & AFPC. hal-04623007.

[23] Cécile Macaire, Chloé Dion, Didier Schwab, Benjamin Lecouteux, and Emmanuelle Esperança-Rodier. Towards Speech-to-Pictograms Translation. In *Interspeech 2024*, pages 857–861, Kos / Greece, 2024. ISCA. hal-04687483.

[24] Cécile Macaire, Emmanuelle Esperança-Rodier, Benjamin Lecouteux, and Didier Schwab. Voice2Picto : un système de traduction automatique de la parole vers des pictogrammes. In *18e Conférence en Recherche d'Information et Applications, 16e Rencontres Jeunes Chercheurs en RI, 30e Conférence sur le Traitement Automatique des Langues Naturelles, 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 10–13, Paris, 2023. ATALA. hal-04208583.

[25] Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. Automatic Speech Recognition and Query By Example for Creole Languages Documentation. In *Findings of the Association for Computational Linguistics : ACL 2022*, Dublin, 2022. hal-03625303.

[26] Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once : is mechanistic interpretability identifiable ? In *The Thirteenth International Conference on Learning Representations*, Singapore, 2025. hal-05226236.

[27] Lucía Ormaechea, Pierrette Bouillon, Maximin Coavoux, Johanna Gerlach, Jérôme Goulian, Benjamin Lecouteux, Cécile Macaire, Jonathan Mutal, Magali Norré, Adrien Pupier, and Didier Schwab. PROPICTO : Developing Speech-to-Pictograph Translation Systems to Enhance Communication Accessibility. In *European Association for Machine Translation*, Tampere, 2023. hal-04241457.

[28] Lucía Ormaechea, Pierrette Bouillon, Johanna Gerlach, Benjamin Lecouteux, Didier Schwab, and Hervé Spechbach. Reconnaissance vocale du discours spontané pour le domaine médical. In *Technologies du Langage Humain (TLH)*, Paris, 2021. hal-03157085.

[29] Lucía Ormaechea, Nikos Tsourakis, Didier Schwab, Pierrette Bouillon, and Benjamin Lecouteux. Simple, Simpler and Beyond : A Fine-Tuning BERT-Based Approach to Enhance Sentence Complexity Assessment for Text Simplification. In *ICNLSP (International Conference on Natural Language and Speech Processing)*, Trento, 2023. hal-04359942.

[30] Lucia Ormaechea-Grijalba, Nikolaos Tsourakis, Pierrette Bouillon, Benjamin Lecouteux, and Didier Schwab. Towards high-quality LLM-based data for French spontaneous speech simplification : an exo-refinement approach. In *Interspeech*, pages 4038–4042, Rotterdam, 2025. hal-05074816.

[31] Lucia Ormaechea-Grijalba, Nikos Tsourakis, Didier Schwab, Pierrette Bouillon, and Benjamin Lecouteux. Simplification Strategies in French Spontaneous Speech. In *DeTermIt ! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, Torino, Italy, 2024. hal-04731828.

[32] Titouan Parcollet, Ha Nguyen, Solène Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Estève, Mickael Rouvier, Jérôme Goulian, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. LeBenchmark 2.0 : A standardized, replicable and enhanced framework for self-supervised representations of French speech. *Computer Speech and Language*, 86 :101622, 2024. hal-04441389.

[33] Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, and Jérôme Goulian. End-to-End Dependency Parsing of Spoken French. In *Interspeech*, Incheon, 2022. hal-03713551.

[34] Vincent Segonne, Aidan Mannion, Laura Cristina Alonzo Canul, Alexandre Audibert, Xingyu Liu, Cécile Macaire, Adrien Pupier, Yongxin Zhou, Mathilde Aguiar, Felix Herron, Magali

Norré, Massih-Reza Amini, Pierrette Bouillon, Iris Eshkol-Taravella, Emmanuelle Esperança-Rodier, Thomas François, Lorraine Goeuriot, Jérôme Goulian, Mathieu Lafourcade, Benjamin Lecouteux, François Portet, Fabien Ringeval, Vincent Vandeghinste, Maximin Coavoux, Marco Dinarelli, and Didier Schwab. Jargon : A Suite of Language Models and Evaluation Tasks for French Specialized Domains. In *LREC-COLING 2024 - Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 9463–9476, Turin, 2024. hal-04535557.

[35] Gilles Sérasset. Towards Sense to Sense Linking across DBnary Languages. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 318–327, Naples, 2025. hal-05267573.

[36] Nakanyseth Vuth, Emrick Poncet, Gilles Sérasset, Didier Schwab, Caroline Djambian, and Benjamin Lecouteux. Ontology-Guided Synthetic Data Generation for Low-Resource Information Extraction : A Case Study in IT Heritage Domain. In *Knowledge Graphs and Large Language Models Workshop (KG-LLM) @ LREC26*, pages 73–81, Palma de Majorque, 2026. hal-05672569.

[37] Nakanyseth Vuth, Gilles Sérasset, and Didier Schwab. KGAST : From Knowledge Graphs to Annotated Synthetic Texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, Bangkok, 2024. hal-04708092.

[38] Yuxuan Zhang, Hippolyte Fournier, Ruslan Kalitvianski, Marco Dinarelli, and Fabien Ringeval. Beyond Static Emotions : Leveraging Multitask Learning to Model Dynamics of Dimensional Affect in Speech. In *28th International Conference on Text, Speech and Dialogue*, volume 16029, pages 109–120, Erlangen-Nürnberg, 2025. Springer Nature Switzerland. hal-05375921.

[39] Yongxin Zhou, François Portet, and Fabien Ringeval. Effectiveness of French Language Models on Abstractive Dialogue Summarization Task. In *LREC 2022*, Marseille, 2022. hal-03727214.

AfIA
Association française
pour l'Intelligence Artificielle

LIS/TALEP : Traitement automatique du langage écrit et parlé

LIS UMR 7020
CNRS et Aix Marseille Université
https://www.lis-lab.fr/talep/

Frederic BECHET
frederic.bechet@univ-amu.fr

Présentation générale

TALEP (Traitement Automatique du Langage Écrit et Parlé) est une équipe de recherche du laboratoire LIS qui développe des méthodes, modèles et ressources pour l'analyse, la compréhension et la génération automatique du langage humain. Ses recherches couvrent l'ensemble du spectre du traitement automatique des langues, depuis la modélisation linguistique fondamentale jusqu'aux applications à fort impact sociétal. L'équipe s'intéresse notamment à l'analyse syntaxique, sémantique et pragmatique des textes et de la parole, aux modèles d'apprentissage profond, aux représentations linguistiques et à leur interprétabilité. Elle explore également les liens entre intelligence artificielle, sciences cognitives et acquisition du langage. Les travaux de TALEP trouvent des applications dans des domaines variés tels que la santé, les humanités numériques, l'éducation, la veille scientifique, les assistants conversationnels et l'interaction humain-machine. Reconnue pour son expertise internationale, l'équipe participe à de nombreux projets nationaux et européens et contribue activement au développement de ressources et logiciels ouverts pour la communauté scientifique.

Positionnement La recherche publique en TAL ne peut rivaliser avec les grands acteurs industriels en termes de moyens techniques. Sa valeur ajoutée réside dans la recherche fondamentale, l'exploration de pistes de recherche n'ayant pas d'applications commerciales immédiates, l'évaluation rigoureuse et indépendante des modèles, ainsi que la prise en compte des impacts sociétaux et environnementaux dans la conception et l'analyse d'impact des méthodes étudiées. L'équipe TALEP se distingue par le couplage entre d'un côté son expertise sur l'analyse et la production de langage par les machines, et d'un autre côté par son articulation étroite avec les sciences cognitives, notamment concernant le développement du langage chez l'enfant, et les sciences humaines et sociales, en particulier la linguistique. Notre projet consiste à affirmer notre rôle sur les fondamentaux du domaine (modélisation, concepts, normes, évaluation) et rappeler que les données langagières sont spécifiques et requièrent des méthodes adaptées tout en plaçant l'humain au centre du processus.

Les questions de recherche prioritaires pour l'équipe porte sur trois axes : (1) **Modèles** : explicabilité et interprétabilité des modèles, frugalité au niveau des ressources linguistiques et computationnelles, élaboration de nouvelles métriques d'évaluation, hybridation entre des modèles de différentes nature ; (2) **Ancrage** : "*confrontation*" entre des données langagières "*pures*" (écrit, parole) et des données "*décrivant le monde*" que ce soit des données sonores, visuelles, des capteurs physique ou biométriques, des signaux cérébraux, problématique de l'alignement des modèles sur des processus cognitifs ou physique à des fins de description et d'automatisation ; (3) **Impact** : l'étude de l'impact des modèles de TAL, sur les interactions sociales, sur des problèmatiques de sécurité, sur les compétences des utilisateurs, tout en ayant soin de toujours considérer l'utilisateur humain comme le centre de nos études.

Principaux thèmes de recherche

Recherche fondamentale en traitement automatique des langues

L'équipe mène des recherches sur les fondements linguistiques du traitement automatique des langues. Elle développe des modèles et des ressources pour l'analyse de textes et de données orales, couvrant notamment la syntaxe, la sémantique, la coréférence, les expressions polylexicales et les phénomènes pragmatiques. Ces travaux contri-

buent à une meilleure représentation des connaissances linguistiques dans les systèmes d'intelligence artificielle. Publications récentes : [8, 11, 14, 10].

Intelligence artificielle et modèles de langage

TALEP conçoit et analyse des modèles d'apprentissage automatique pour le langage, en particulier des approches fondées sur l'apprentissage profond. Les recherches portent sur les représentations linguistiques apprises par les modèles, leur capacité de généralisation, leur robustesse, la capacité à rendre ces modèles frugaux en termes de calcul et de données d'apprentissage et leur explicabilité. L'équipe s'intéresse aussi fortement aux liens entre ces modèles et les mécanismes cognitifs impliqués dans l'acquisition et le traitement du langage. Publications récentes : [3, 7, 12, 20]

Langage, interaction et multimodalité

L'un des axes majeurs de l'équipe concerne la modélisation du langage en interaction. Les chercheurs étudient la manière dont les informations verbales, visuelles et comportementales se combinent dans les échanges humains. Ces travaux alimentent le développement d'agents conversationnels, d'avatars animés et de robots sociaux capables d'interagir de manière naturelle avec leurs utilisateurs. Publications récentes : [1, 16, 5, 15]

Applications sociétales

Les technologies développées par TALEP sont appliquées à de nombreux domaines. En santé, elles contribuent à l'analyse documentaire, à la veille scientifique et à l'exploitation de données textuelles. Dans les humanités numériques, elles permettent l'exploration de corpus historiques et d'archives. L'équipe développe également des solutions pour l'éducation, les assistants numériques, la recherche d'information et l'analyse de documents techniques. Publications récentes : [4, 17, 13, 2]

Ressources, logiciels et science ouverte

TALEP participe activement à la création et à la diffusion de ressources linguistiques, corpus annotés et logiciels open source. Ces ressources couvrent aussi bien la langue écrite que la langue parlée et sont largement utilisées par la communauté scientifique internationale. Cette politique de science ouverte contribue à la reproductibilité des recherches et à l'émergence de nouvelles collaborations. Publications récentes : [18, 19, 6, 9]

Membres Impliqués

- Marianna APIDIANAKI (CR, CNRS)
- Leonor BECERRA (MCF HDR, AMU)
- Frédéric BECHET (PR, AMU)
- José DELOFEU (PR, AMU)
- Benoit FAVRE (PR AMU, en détachement)
- Abdellah FOURTASSI (MCF HDR, AMU)
- Marvin LAVECHIN (CR CNRS)
- Alexis NASR (PR, AMU)
- Magalie OCHS (PR, AMU)
- Carlos RAMISCH (MCF HDR, AMU)
- Marie-Hélène STEFANINI (MCF, AMU)
- Michael ZOCK (DR, CNRS, émérite)

Références

[1] Abhishek Agrawal, Benoit Favre, and Abdellah Fourtassi. Communicative Intent Coordination in Child-Caregiver Interactions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Rotterdam (NL), Netherlands, July 2024.

[2] Eunice Akani, Frédéric Bechet, Benoît Favre, and Romain Gemignani. Unified Framework for Spoken Language Understanding and Summarization in Task-Based Human Dialog processing. In *INTERSPEECH 2024*, Kos Island, Greece, September 2024.

[3] Elie Antoine, Frédéric Béchet, Géraldine Damnati, and Philippe Langlais. A linguistically-motivated evaluation methodology for unraveling model's abilities in reading comprehension tasks. In *Proceedings of EMNLP 2024*, pages 18376–18392, Miami, USA, November 2024. ACL.

[4] Ikram Belmadani, Oumaima El Khettari, Pacôme Constant Dit Beaufils, Benoit Favre, and Richard Dufour. MedInjection-FR : Exploring the Role of Native, Synthetic, and Translated Data in Biomedical Instruction Tuning. In

LREC 20, Majorqua, Spain, May 2026. ELRA Language Resources Association.

[5] Alice Delbosc, Magalie Ochs, Nicolas Sabouret, Brian Ravenet, and Stéphane Ayache. Mitigation of gender bias in automatic facial non-verbal behaviors generation. In *ICMI '24*, pages 284–292, San Jose Costa Rica, France, November 2024. ACM.

[6] Goumri Dhia-Elhak, Agrawal Abhishek, Nikolaus Mitja, Hong Duc Thang Vu, Bodur Kübra, Semmar Elias, Armand Cassandre, Mazzocconi Chiara, Gupta Shreejata, Laurent Prevot, Favre Benoit, Becerra-Bonache Leonor, and Fourtassi Abdellah. CHICA : A Developmental Corpus of Child-Caregiver's Face-to-face vs. Video Call Conversations in Middle Childhood. In *LREC-COLING*, Torino, Italy, September 2024.

[7] Loïc Fosse, Frédéric Béchet, Benoît Favre, Géraldine Damnati, Gwénolé Lecorvé, Maxime Darrin, Philippe Formont, and Pablo Piantanida. Statistical Deficiency for Task Inclusion Estimation. In *Proceedings of the 63rd Annual Meeting of the ACL (Volume 1 : Long Papers)*, pages 382–415, Vienna, France, July 2025. ACL.

[8] Ioana Ivan, Carlos Ramisch, and Alexis Nasr. Evaluating Pretrained Causal Language Models for Synonymy. In *Findings of the Association for Computational Linguistics : ACL 2025*, pages 12533–12551, Vienna, France, July 2025. Association for Computational Linguistics.

[9] Hichem Ammar Khodja, Frédéric Béchet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. WikiFactDiff : A Large, Realistic, and Temporally Adaptable Dataset for Atomic Factual Knowledge Update in Causal Language Models. In *Proceedings of the 2024 Joint International Conference LREC-COLING*, pages 17614–17624, Turin, Italy, May 2024.

[10] Marvin Lavechin and Thomas Hueber. From perception to production : how acoustic invariance facilitates articulatory learning in a self-supervised vocal imitation model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23863–23874, 2025.

[11] Anna Mosolova, Marie Candito, and Carlos Ramisch. In the LLM era, Word Sense Induction remains unsolved. Findings of the ACL 2025, pages 17161–17178, Vienne, Austria, July 2025. ACL.

[12] Mitja Nikolaus and Abdellah Fourtassi. Modelling children's grammar learning via caregiver feedback in natural conversations. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 381(1943), February 2026.

[13] Magalie Ochs, Ristorcelli Marion, D Ambra, Vessiot Manon, Rémy Casanova, and Jean-Marie Pergandi. REVITALISE viRtual skIlls TrAining for pubLIc SpEaking. In *Intelligent Virtual Agent*, Glasgow, United Kingdom, September 2024.

[14] Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. StyleDistance : Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of 2025 NAACL (Volume 1 : Long Papers)*, pages 8662–8685, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[15] Morgane Peirolo, Zihan Xu, and Abdellah Fourtassi. Development of Flexible Role-Taking in Conversations Across Preschool. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Rotterdam (NL), Netherlands, July 2024.

[16] Marion Ristorcelli, Elodie Etienne, Michaël Schyns, Rémy Casanova, and Magalie Ochs. Evaluating Multimodal Behavioral Features for Public Speaking Assessment in Virtual Reality. In *ACM International Conference on Intelligent Virtual Agents (IVA '25)*, number 40, pages 1–4, Berlin, Germany, September 2025. Conference - Intelligent Virtual Agents, ACM.

[17] Ricardo Rodriguez, Stéphane Huet, Benoît Favre, and Mickael Rouvier. Comparative Analysis of Human and Large Language Model Performance in Pharmacology Multiple-

Choice Questions. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*, pages 1022–1029, Varna, Bulgaria, September 2025. ACL.

[18] Manon Scholivet, Agata Savary, Louis Estève, Marie Candito, and Carlos Ramisch. SELEXINI - a large and diverse automatically parsed corpus of French. In *BUCC - Building and Using Comparable Corpora*, Abu Dhabi, United Arab Emirates, January 2025.

[19] Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Mitrofan, and Vasile Pais. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 254–275, Rabat, Morocco, March 2026. Association for Computational Linguistics.

[20] Michael Zock. The mental lexicon : A blueprint for the dictionaries of tomorrow ? *Frontiers in Artificial Intelligence*, 5 :4, January 2023.

AfIA
Association française
pour l'Intelligence Artificielle

LISN/STL : Sciences et Technologies des Langues

LISN / LIPS
https://www.lisn.upsaclay.fr/recherche/departements-et-equipes/sciences-et-technologies-des-langues/lips/

LISN / M³
https://www.lisn.upsaclay.fr/recherche/departements-et-equipes/sciences-et-technologies-des-langues/m3/

LISN / SEME
https://www.lisn.upsaclay.fr/recherche/departements-et-equipes/sciences-et-technologies-des-langues/seme/

Ioana VASILESCU
Responsable de l'équipe LIPS
ioana.vasilescu@lisn.fr

Albert RILLIARD
Responsable de l'équipe M³
albert.rilliard@lisn.fr

Agata SAVARY
Responsable de l'équipe SEME
agata.savary@lisn.fr

Introduction

Le département Sciences et Technologies de la Langue (STL) regroupe les activités du LISN en **traitement des langues** au sein des équipes LIPS, M3 et SEME. Le spectre des recherches conduites dans ces trois équipes est très large et s'étend à tous les aspects de la langue, **parlée, écrite ou signée**.

L'équipe LIPS (Langue Interaction Parole et Signes) s'intéresse à la description et la modélisation interdisciplinaires de la langue sous toutes ses formes (orale, écrite ou signée). L'équipe mobilise des méthodes d'analyse linguistique et de traitement de la parole pour réaliser des recherches à la croisée des sciences numériques et des sciences humaines et sociales. L'équipe propose des contributions à l'extraction d'informations dans les dialogues, la modélisation et traitement des langues des signes, le traitement de la parole et la modélisation des variations multilingues.

L'équipe M3 (Modèles, Méthodes, et Multilinguisme) s'intéresse au développement de modèles et méthodes pour faciliter la découverte des propriétés fondamentales du langage. Pour cela, l'équipe s'appuie sur des approches linguistiques et statistiques — qu'elles soient computationnelles (modèles neuronaux, méthodes stochastiques ou symboliques), linguistiques (typologie des langues, diversité linguistique, et universaux, affects) ou sociétales (accessibilité, incitations, préservation des langues, traitement des langues et dialectes défavorisés).

L'équipe SEME (Sémantique et Extraction d'Information) s'intéresse à l'extraction du sens des énoncés à des fins d'analyse, de compréhension, de modélisation ou de production. L'équipe développe des approches linguistiques, statistiques ou neuronales pour l'analyse des documents écrits notamment dans des domaines de spécialité comme la santé ou l'astrophysique. L'équipe propose des contributions pour l'extraction d'informations, la sémantique lexicale (expressions polylexicales, ...) et l'évaluation des impacts environnementaux et sociétaux des technologies de la langue.

Nous présentons ci-dessous des thèmes de recherche transverses impliquant des membres de plusieurs équipes du département.

Thèmes de recherche

Modélisation et traitement automatique des langues des signes (LIPS)

Les Langues des Signes (LS) sont des langues naturelles pratiquées au sein des communautés de Sourds et la Langue des Signes Française (LSF) est celle utilisée en France. Ce sont des langues visuo-gestuelles : une personne s'exprime en LS en utilisant de nombreuses composantes corporelles (les mains et les bras, mais aussi différentes du visage, le regard, le buste, etc.) et son interlocuteur perçoit le

message par le canal visuel. Le système linguistique des LS exploite ces canaux spécifiques : de nombreuses informations sont exprimées simultanément et s'organisent dans l'espace, et l'iconicité joue un rôle central.

À ce jour, les LS n'ont pas de système d'écriture ni de système graphique standard pour la transcription et sont encore peu décrites et peu dotées. La modélisation et le traitement automatique des LS (TALS), domaine de recherche récent, nécessitent de concevoir des représentations dans un domaine où peu de données sont disponibles et où les modèles pré-existants, essentiellement linéaires, ont été développés pour les langues écrites ou parlées et ne couvrent pas tous les aspects des LS. Les tâches de TALS que nous étudions concernent la reconnaissance [19], la génération et la traduction, automatiques ou assistées, en s'appuyant sur des corpus et ensembles de données que nous produisions ou issus de médias, ainsi que des modèles linguistiques [5].

Variation, diversité et universaux linguistiques (LIPS, M^3, SEME)

Au sein de cette thématique, nous nous intéressons à deux notions connexes : la variation et la diversité linguistiques. Nous observons la variation linguistique au sein d'une langue et effectuons sa conceptualisation, ainsi que son analyse qualitative. Dans des contextes multilingues, nous analysons des énoncés à l'aide de formalismes unifiés à travers de nombreuses langues, afin de questionner l'existence d'universaux linguistiques. Nous cherchons également à conceptualiser la quantification de la diversité linguistique pour les besoins du traitement des langues.

Dans ces cadres, nous nous intéressons notamment aux questions de recherche suivantes :

- Comment relier la variation diachronique et synchronique [4, 7] ?
- Quelles stratégies expressives, universelles ou spécifiques, s'observent lors d'interactions en face à face [24] ?
- Avec quelle granularité géographique et quels traits phonétiques peut-on caractériser les variations des accents et des dialectes [3] ?
- Comment définir l'effort vocal en termes de variation prosodique [23] ?
- Comment unifier l'annotation de textes aux niveaux morphologique, syntaxique et sémantique afin de faire émerger des "universaux statistiques" [18, 26, 17, 25, 14] ?
- Comment définir des mesures de distance entre objets linguistiques (entre phones, syllabes, mots, phrases, langues, etc.) et entre leurs distributions [9] ?
- Comment théoriser et quantifier la diversité linguistique dans le domaine du traitement des langues [12, 27] ?

Apprentissage automatique, modèles et algorithme (LIPS, M^3, SEME)

Cette thématique regroupe les activités du département STL dans la conception de méthodes générales pour l'apprentissage automatique. Notre recherche porte, entre autres, sur les méthodes neuronales, les approches de modélisation Bayésienne, les problèmes d'algorithmie, de frugalité et leur évaluation. Cette thématique trouve des applications sur des modalités diverses (parole, image, vidéo et autres).

Les méthodes sont développées afin de traiter et de modéliser de multiples modalités, telles que des langues naturelles (écriture, parole, signes, etc...), l'image et la vidéo. Cette thématique est transverse à toutes les équipes de STL et vise aussi bien des applications de traitement des langues que l'étude de problématiques linguistiques [28, 10, 15].

Corpus, Représentations et Evaluation (LIPS, M3, SEME)

Ce thème traite de l'aspect théorique et des problèmes pratiques concernant la collecte, l'annotation et la diffusion de grands corpus multilingues. Il aborde aussi la question des méthodes d'évaluation dans un contexte d'émergence d'outils multitâches comme les LLMs. Pour cela, il faut collecter des faits de langues, des écrits, des enregistrements de sons ou de gestes, selon une hypothèse de recherche. Il faut ensuite élaborer des représentations des structures linguistiques émergentes, allant des syllabes, caractères ou gestes, aux interprétations

de plus haut niveau, objectives ou subjectives et relevant de la sémantique ou de la pragmatique. Ces représentations servent ensuite pour annoter les structures en contexte. Les corpus ainsi annotés peuvent être utilisés pour la documentation des langues, l'enseignement et l'évaluation des technologies en comparant les performances des machines à celles des humains.

Les questions de recherche associées à cette thématique sont :

- Comment définir les critères prévalents pour constituer un corpus en fonction d'une tâche, une langue et un média donné ?
- Quelle représentation utiliser pour expliciter les phénomènes interprétatifs ou génératifs que l'on souhaite modéliser [21] ?
- Comment prendre en compte la subjectivité ou les biais culturels dans l'annotation [6] ?
- Quels logiciels sont nécessaires pour la tâche d'annotation humaine ?
- Comment garantir un bon accord inter-annotateurs ?
- Comment estimer la qualité d'un corpus et d'une annotation par rapport à une tâche donnée ?
- Comment évaluer la performance d'un algorithme d'analyse ou de génération ?
- Est-il possible de bénéficier des progrès apportés par les grands modèles de langue sans dénaturer la tâche d'annotation [1] ?

Traitement Automatique des Langues appliqué au domaine biomédical (SEME)

L'équipe SEME développe et adapte des méthodes et ressources en traitement automatique des langues (TAL) pour le domaine médical. Elle mène des recherches sur les textes cliniques (compte-rendus hospitaliers) et les textes contenant des informations cliniques (littérature scientifique, réseaux sociaux, forums de santé) par les tâches habituelles du TAL (extraction d'information, classification de documents, génération de documents) avec des objectifs de santé publique (pharmacovigilance, études rétrospectives...). L'équipe travaille à modéliser la langue biomédicale, soit pour annoter des corpus, soit pour développer des modèles de langue adaptés au domaine biomédical conciliant robustesse et sécurité des informations. Elle entretient des liens avec plusieurs acteurs de la santé (AP-HP, CHU de Bordeaux, Rouen, Marseille, Health Data Hub, Santé Publique France, etc.).

Les activités de recherche de l'équipe s'articulent autour de plusieurs axes complémentaires :

- **Extraction d'information et compréhension de textes cliniques.** Nous développons des méthodes pour extraire et structurer l'information contenue dans les textes médicaux, notamment les entités, relations et événements cliniques. Ces travaux visent à transformer des documents non structurés en représentations exploitables pour l'analyse médicale et la recherche clinique [2, 13].
- **Ressources, évaluation et données synthétiques.** Un axe important concerne la construction de ressources linguistiques, la création de corpus d'évaluation et l'étude de données synthétiques pour le domaine clinique. Ces travaux permettent de mieux évaluer les systèmes et d'explorer de nouvelles méthodes d'entraînement dans des contextes sensibles [1, 11].
- **TAL pour les réseaux sociaux et la santé publique.** Nous appliquons des méthodes de TAL à l'analyse des réseaux sociaux et des données générées par les utilisateurs afin de détecter des signaux liés à la santé, notamment dans le cadre de la surveillance des effets indésirables de médicaments [22].

Impacts environnementaux et sociétaux de l'IA (SEME, M^3)

L'utilisation généralisée des outils fondés sur des modèles d'IA et en particulier de traitement automatique des langues, pose question du fait des impacts sociaux et environnementaux engendrés par les infrastructures d'IA. En effet la fabrication, utilisation et fin de vie des équipements ont des conséquences sur les travailleurs impliqués, et de nombreuses pollutions, que ce soit en termes d'empreinte carbone, de consommation d'eau ou d'épuisement des ressources par exemple. En outre, l'IA peut contribuer à améliorer ou détériorer les impacts d'autres secteurs, par exemple en réduisant la consommation énergétique des bâtiments, ou inversement en incitant à la surconsommation de produits ou de vidéos. Cette théma-

tique vise à évaluer ces impacts, les analyser, et à proposer des leviers de réduction de l'empreinte environnementale[20, 16]. La thématique s'intéresse également à caractériser les biais, notamment stéréotypés, véhiculés par les modèles d'IA. Il s'agit ainsi de proposer des méthodes d'évaluation et de mitigation des biais et de leurs impacts, en lien avec les sciences sociales [8].

Membres permanents

LIPS

- Annelies Braffort (CNRS)
- Michael Filhol (CNRS)
- Jean-Luc Gauvain (Emerite, CNRS)
- Sahar Ghannay (Université Paris-Saclay)
- Michèle Gouiffès (Université Paris-Saclay)
- Camille Guinaudeau (Université Paris-Saclay)
- Lori Lame (Emerite, CNRS)
- Sophie Rosset (CNRS)
- Ioana Vasilescu (CNRS)

M^3

- Gilles Adda (CNRS)
- Philippe Boula De Mareüil (CNRS)
- Laurence Devillers (Univ. Paris-Sorbonne)
- Marc Evrard (Université Paris-Saclay)
- Kim Gerdes (Université Paris-Saclay)
- Thomas Lavergne (Université Paris-Saclay)
- Jean-Sylvain Liénard (Émérite, CNRS)
- Joseph-Jean Mariani (Émérite, CNRS)
- Lucas Ondel Yang (CNRS)
- Hélène Maynard (Université Paris-Saclay)
- Albert Rilliard (CNRS)

SEME

- Eva Feillet (Université Paris-Saclay)
- Thomas Gerald (Université Paris-Saclay)
- Cyril Grouin (CNRS)
- Thierry Hamon (Université Paris-Nord)
- Gabriel Illouz (Université Paris-Saclay)
- Anne-Laure Ligozat (ENSIIE)
- Nona Naderi (Université Paris-Saclay)
- Takuya Nakamura (CNRS)
- Aurélie Névéol (CNRS)
- Patrick Paroubek (CNRS)
- Agata Savary (Université Paris-Saclay)
- Anne Vilnat (Emerite, Université Paris-Saclay)
- Pierre Zweigenbaum (Emerite, CNRS)

Références

[1] Mathilde Aguiar, Pierre Zweigenbaum, and Nona Naderi. Assessing the Difficulty of Inference Types in Natural Language Inference for Clinical Trials. In *The Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 5290–5300, Palma, France, May 2026.

[2] Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier, and Aurélie Névéol. Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics*, 130 :104073, 2022.

[3] Philippe Boula de Mareüil and Paolo Roseano. A speaking atlas of the languages of the Iberian Peninsula : focus on rhythm and varieties in contact. *Dialectologia*, 35 :27–54, July 2025.

[4] Philippe Boula de Mareüil, Marc Evrard, Alexandre François, and Antonio Romano. Computer modelling of innovations relative to latin in contemporary romance dialects. *Isogloss. Open Journal of Romance Linguistics*, 11(3) :1–31, Mar. 2025.

[5] Annelies Braffort, Michael Filhol, Michèle Gouiffès, Julie Halbout, and Julie Lascar. Sign Language Processing with Linguistic Structure. BMVA Symposium on AI for Sign Language Translation, Production, and Linguistics, December 2025. Poster.

[6] Gustave Cortal and Alain Finkel. Formalizing style in personal narratives. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7311–7326, Suzhou, China, November 2025. Association for Computational Linguistics.

[7] Simon Devauchelle, David Doukhan, Rémi Uro, Lucas Ondel Yang, Valentin Pelloin, Olympia Imbert-Brégégère, Véronique Lefort, Kévin Picard, Emeline Seignobos, and Albert Rilliard. spinach : A diachronic corpus of french broadcast speech controlled for speakers' age and gender. In *Fifteenth International Conference on Language Resources and Evaluation*

(LREC 2026), Palma, Mallorca, Spain, 2026. European Language Resources Association.

[8] Fanny Ducel, Aurélie Névéol, and Karën Fort. " you'll be a nurse, my son ! " automatically assessing gender biases in autoregressive language models in french and italian. *Language Resources and Evaluation*, pages 1495–1523, 2024.

[9] Louis Estève, Agata Savary, and Thomas Lavergne. Vector spaces for quantifying disparity of multiword expressions in annotated text. In Xiyan Fu and Eve Fleisig, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4 : Student Research Workshop)*, pages 110–130, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[10] Louis Estève, Christophe Servan, Thomas Lavergne, and Agata Savary. A diversity diet for a healthier model : A case study of French ModernBERT. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Findings of the Association for Computational Linguistics : ACL 2026*, pages 34168–34181, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[11] Foucauld Estignard, Sahar Ghannay, Julien Girard-Satabin, Nicolas Hiebel, and Aurélie Névéol. Evaluating the confidentiality of synthetic clinical texts generated by language models. In *International Conference on Artificial Intelligence in Medicine*, pages 130–139. Springer, 2025.

[12] Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. A survey of diversity quantification in natural language processing : The why, what, where and how, 2026.

[13] Thibaut Fabacher, Erik-André Sauleau, Emmanuelle Arcay, Bineta Faye, Maxime Alter, Archia Chahard, Nathan Miraillet, Adrien Coulet, and Aurélie Névéol. Efficient extraction of medication information from clinical notes : an evaluation in 2 languages. *Journal of the American Medical Informatics Association*, 32(12) :1855–1864, 12 2025.

[14] Pegah Faghiri, Kim Gerdes, and Sylvain Kahane. Verifying the menzerath-altmann law in the verbal domain in 180 languages. In *Proceedings of the Ninth Workshop on Universal Dependencies (UDW 2026)*, pages 76–90, Palma di Mallorca, Spain, 2026. ELRA.

[15] Eva Feillet, Ryan Whetten, David Picard, and Alexandre Allauzen. Polynomial mixing for efficient self-supervised speech encoders. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1401–1405, 2026.

[16] Marion Ficher, Tom Bauer, and Anne-Laure Ligozat. A comprehensive review of the end-of-life modeling in lcas of digital equipment. *The International Journal of Life Cycle Assessment*, 30(1) :20–42, 2025.

[17] Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, and Yixuan Li. Joint annotation of morphology and syntax in dependency treebanks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9568–9577, Torino, Italia, May 2024. ELRA and ICCL.

[18] Sylvain Kahane, Ziqian Peng, and Kim Gerdes. Word order flexibility : a typometric study. In Owen Rambow and François Lareau, editors, *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)*, pages 68–80, Washington, D.C., March 2023. Association for Computational Linguistics.

[19] Julie Lascar, Jules Françoise, Michèle Gouiffès, Annelies Braffort, and Diandra Fabre. PoET : Lightweight Pose Encoder Transformer for Online Sign Language Recognition. In Antonino Furnari and Petia Radeva, editors, *Proceedings of the 21st International Conference on Computer Vision Theory and Applications*, volume 1, pages 19–28, Marbella, Spain, March 2026. SCITEPRESS - Science and Technology Publications.

[20] Clément Morand, Anne-Laure Ligozat, and Aurélie Névéol. Mlca : a tool for machine learning life cycle assessment. In *2024 10th International Conference on ICT for Sustainability (ICT4S)*, pages 227–238. IEEE, 2024.

[21] Luc Pommeret, Thomas Gerald, Christophe Servan, Sahar Ghannay, Patrick Paroubek, and Sophie Rosset. Étude des propositionneurs multilingues : évaluation et interprétabilité. In *Actes de la 33+.5exème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Nantes, June 2026. ATALA.

[22] Lisa Raithel, Philippe Thomas, Roland Roller, Oliver Sapina, Sebastian Möller, and Pierre Zweigenbaum. Cross-lingual approaches for the detection of adverse drug reactions in German from a patient's perspective. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3637–3649, Marseille, France, June 2022. European Language Resources Association.

[23] Albert Rilliard, Christophe d'Alessandro, and Marc Evrard. Paradigmatic variation of vowels in expressive speech : Acoustic description and dimensional analysis. *The Journal of the Acoustical Society of America*, 143(1) :109–122, 01 2018.

[24] Albert Rilliard, João Antônio De Moraes, Donna Erickson, Marine Guerry, Angelika Hönemann, Tan Lee, Matthew King-Hang Ma, Hansjörg Mixdorff, Preeti Rao, and Takaaki Shochi. Cross-cultural dimensions organizing prosodic attitudes reception. *Journal of Speech Sciences*, 14 :e025012, October 2025.

[25] Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Pais, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, Chaya Liebeskind, Irina Lobzhanidze, Aleksandra M. Marković, Gunta Nešpore-Bērzkalne, Adriana S. Pagano, Mehrnoush Shamsfard, Ranka Stankovic, Vahide Tajalli, Carole Tiberius, and Aakanksha Padhye. Parseme 2.0 multilingual corpus of multiword expressions. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral, editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4819–4834, Palma, Mallorca, Spain, May 2026. European Language Resources Association (ELRA).

[26] Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. PARSEME meets Universal Dependencies : Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9, 2023.

[27] Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Mitrofan, and Vasile Pais. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In Atul Kr. Ojha, Verginica Barbu Mititelu, Mathieu Constant, Ivelina Stoyanova, A. Seza Doğruöz, and Alexandre Rademaker, editors, *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 254–275, Rabat, Marocco, March 2026. Association for Computational Linguistics.

[28] Lucas Ondel Yang, Tina Raissi, Martin Kocour, Pablo Riera, and Caio Corro. Fast and general automatic differentiation for finite-state methods. *arXiv preprint arXiv :2602.12300*, 2026.

AfIA
Association française
pour l'Intelligence Artificielle

LIUM/LST : Language and Speech Technologies

LIUM - LST EA4023
Le Mans Université
lium.univ-lemans.fr

Marie TAHON
marie.tahon@univ-lemans.fr

Présentation générale

Le LIUM (Laboratoire d'Informatique de l'Université du Mans) est un laboratoire d'informatique composé d'une trentaine d'enseignant.es-chercheur.es. Il est composé de deux équipes :

- EIAH : Environnement Informatique pour l'Apprentissage Humain
- LST : Language and Speech Technologies

Les recherches de l'équipe LST s'inscrivent dans le traitement automatique des langues orales et écrites. L'équipe a fait partie des pionniers dans l'utilisation du paradigme neuronal pour l'apprentissage de modèles de reconnaissance automatique de la parole. Elle a une expertise historique à la fois sur les données (collecte et annotation) et sur la résolution de différentes tâches associées au traitement automatique du signal de parole et du texte écrit qui sont détaillées ci-dessous. Sa particularité est de pouvoir associer des connaissances fortes sur les modèles de traitement automatique avec une analyse poussée du signal de parole ou du contenu linguistique. La présence d'une linguiste et d'un statisticien dans l'équipe renforce efficacement cette complémentarité.

Principaux thèmes de recherche

Collecte et annotation de données

L'équipe LST possède une expertise reconnue dans la collection, l'annotation et la distribution de corpus ainsi que dans le développement des protocoles d'évaluation qui leur sont associés. LST a développé des processus d'annotation semi-automatiques permettant d'alléger les coûts de production et de maximiser la couverture des données avec des annotations dédiées à différentes tâches, par exemple sur la segmentation et regroupement en locuteur [28].

Ces dernières années, plusieurs protocoles d'annotation ont été réalisés afin de mieux caractériser certains phénomènes subjectifs dans la parole spontanée. Par exemple, un corpus de plus de 30h de conversations téléphoniques issues de centres d'appel a été annoté sur une dimension affective adaptée au contexte. Cette dimension suit un axe construit pour cette étude : frustration-satisfaction [9]. D'autres protocoles d'annotation ont été mis en œuvre pour la terminalité [32], ou la parole superposée [8] dans des données d'archives radio/TV.

Dans la continuité de ses travaux en traduction du texte et de la parole, l'équipe a également développé une expertise sur la collecte de données langagières dans des langues peu dotées. Ces données peuvent prendre la forme de segmentation temporelle, transcription orthographique et traduction. Parmi les langues étudiées, nous pouvons citer l'Arabe Tunisien [1] et plusieurs variantes de dialectes Kurde [20, 19].

Segmentation du signal acoustique

Le LIUM occupe une place de référence internationale dans le domaine de la segmentation et du regroupement en locuteurs [16, 28]. Ce savoir-faire s'étend à la segmentation et localisation de locuteurs pour le traitement d'antennes de microphones [11] ou aux tâches conjointes de segmentation, par exemple en parole superposée [7] et séparation de sources [24].

Une part récente des travaux de l'équipe LST s'attaque à la détection et segmentation de contrefaçons dans le signal audio [30] principalement, avec des perspectives à court terme sur les signaux vidéos.

Interactions vocales et compréhension de la parole

La compréhension de la parole fait partie des expertises historiques de l'équipe LST. Les travaux se sont initialement concentrés sur des tâches typiquement NLP, comme la détection d'entités nommées, pour aller vers des approches prenant en entrée un signal de parole directement [15], ce qui a permis de développer des modèles neuronaux de bout en bout [23]. Des travaux récents sur le développement de méthodologies robustes pour la compréhension de la parole sur des langues peu dotées (langues africaines notamment) permettent d'étudier des problématiques nouvelles, intégrant des modèles de langages multilingues.

L'étude des interactions vocales humaines, dans des contextes spontanés et expressifs fait partie de l'expertise de l'équipe en traitement de la parole. En particulier, des modèles de détection de satisfaction dans les conversations téléphoniques ont été mis en place, ce qui a permis de mettre en évidence l'importance d'une sémantique de la disfluence, marquée autant à l'oral que dans la transcription [29]. Le projet GEM visant à étudier les interruptions hommes/femmes, a apporté un cadre favorable pour la caractérisation humaine et automatique des interruptions dans les débats médiatiques [8]. De plus, il nous a permis d'étudier le phénomène de terminalité du discours, une notion fondamentale dans la gestion des tours de parole [32].

L'équipe a également été moteur sur le développement de compétences sur les interactions vocales Humain-machine, entre autres pour la mise en place de méthodologies de contrôle prosodique et phonétique pour la synthèse expressive. Des premiers travaux pour le contrôle de la synthèse, s'appuyant sur la manipulation de représentations temporelles intermédiaires de type PhoneticPosteriorGram (PPG) ont mis en évidence le fait que les PPG supprimaient l'information locuteur [5], ce qui a un intérêt pour démêler les aspects locuteurs et phonétiques par la suite.

Par ailleurs, l'équipe développe des méthodes de sélection de prompts informatifs appliquées au clonage de voix [14] en collaboration avec l'entreprise Voxygen. Ces recherches s'articulent autour de l'évaluation et de l'optimisation de la dimension persuasive de la parole synthétisée et s'appuient sur l'apprentissage par renforcement et le feedback.

Sécurité et biométrie vocale

Le LIUM dispose d'une expertise historique dans le domaine de la sécurité vocale, incluant les différentes tâches de reconnaissance du locuteur (segmentation, regroupement, identification, vérification) et maintient une plateforme propriétaire et des modèles à l'état-de-l'art pour toutes ces tâches. L'équipe LST participe depuis de nombreuses années à l'évolution de ce domaine pour l'ensemble de la chaîne de traitement en travaillant dans le cadre de projets collaboratifs tels que DeepPrivacy, sur la pseudo-anonymisation [22], la synthèse [26, 14], la détection de contenus falsifiés audio multilingues [31]. Les travaux de l'équipe intègrent tous des problématiques d'interprétabilité et de biais qui sont inhérentes à ce domaine et se déploient également pour des approches multimodales [13].

Interprétabilité - modèles et évaluation

Le LIUM développe une expertise dans l'apprentissage de représentations interprétables à travers différentes modalités, principalement audio et texte.

Une première direction consiste à concevoir des représentations interprétables à partir de représentations denses et emmêlées issues d'un modèle pré-entraîné. Ces approches ont été appliquées à la segmentation audio en exploitant la factorisation matricielle non-négative [10]. Les derniers travaux menés exploitent les auto-encodeurs parcimonieux notamment pour interpréter des représentations de locuteurs [27] ou de modèles audio auto-supervisés [12].

Une seconde direction est de développer des approches nouvelles qui contraignent l'espace de représentation à encoder des facteurs explicatifs connus à l'aide de prototypes [2] ou en conditionnant l'espace à certains facteurs [3].

L'équipe développe également des modèles de représentations textuelles interprétables, en particulier SINr[4] et DINr, deux modèles basés graphes

4. https://github.com/SINr-Embeddings/sinr/

pour les représentations lexicales interprétables *by design*, par opposition aux représentations interprétables extraites post-hoc.

L'équipe travaille à mettre l'humain au cœur des évaluations de l'interprétabilité des représentations extraites. Ceci a amené le développement de protocoles expérimentaux originaux de tests perceptifs. Semiolex[5], qui est un *Game With a Purpose*, en est un exemple : le but est de collecter de l'information sur la façon optimale de présenter des représentations interprétables. Un test d'intrusion pour évaluer l'interprétabilité de dimensions représentant le locuteur a été proposé [25], et ouvre des perspectives de recherche intéressantes à la fois sur l'interprétabilité et l'apprentissage des modèles eux-mêmes.

Stratégies d'apprentissage centrées sur l'humain, évaluation perceptive

Une partie des travaux de l'équipe a pour objectif d'intégrer l'humain dans la stratégie d'apprentissage des modèles de traitement de la parole et du texte. Cet aspect nécessite de modéliser la perception humain vis à vis d'une tâche spécifique, mais également de développer des méthodes nouvelles d'apprentissage.

Par exemple, certains travaux récents de l'équipe visent à lever le verrou du coût de l'évaluation humaine du naturel de la parole en développant des métriques objectives robustes [17]. L'approche explore le détournement de modèles de détection de contenus falsifiés en prédicteurs de qualité fondés sur la perception humaine [18]. Elle intègre également des stratégies d'apprentissage continu et actif afin de permettre une adaptation trans-lingue et trans-domaine performante des prédicteurs de scores MOS face aux ruptures technologiques.

Face à la vulnérabilité des systèmes aux attaques de synthèse vocale inédites, la problématique est d'optimiser la stabilité et la généralisation des modèles de détection. La solution s'appuie sur des stratégies de Curriculum Learning basées sur des mesures de confiance par échantillon, permettant de réduire drastiquement les temps de calcul tout en améliorant la robustesse temporelle des classifieurs.

L'apprentissage actif et l'apprentissage continu sont également au cœur des préoccupations de l'équipe LST afin de réduire le coût humain et matériel de développement et de maintenance des systèmes automatiques. En plus de systèmes dédiés et de protocoles d'apprentissage, le LIUM s'investit dans l'évaluation de ces approches et la création de ressources spécifiques.

Ressources limitées et Multilinguisme

L'équipe LST a montré sa capacité à développer des méthodes et systèmes dédiés à des cas d'usage où les ressources, puissance de calcul ou quantité de données, sont fortement limitées, et ce, pour différentes tâches.

Dans le cas des langues peu dotées, le LIUM s'est illustré en menant un projet collaboratif de grande ampleur dédié à la traduction parole vers parole [6]. Il participe également régulièrement aux campagnes d'évaluation IWLST dont une des tâches se focalise sur les langues à faibles ressources. Par exemple, le LIUM a proposé en 2026 une tâche d'évaluation sur le Kurde [19]. La distillation de modèles grâce à l'utilisation de pseudo-labels a été montrée comme étant une approche efficace pour l'apprentissage de modèles frugaux, par exemple pour des cas d'usage de la traduction vocale avec des ressources computationnelles limitées [21].

À travers des travaux comme [4], l'équipe LST questionne les paradigmes dominants et propose des architectures neuronales redéfinies pour la contrainte de flux (streaming).

La problématique concerne la forte limitation des données d'apprentissage et de la puissance de calcul pour les langues peu dotées. Les contributions scientifiques se structurent autour de la conception d'architectures de synthèse vocale frugales. Pour pallier le manque de données annotées, la solution intègre le développement de métriques automatiques de sélection de systèmes ainsi que des approches d'apprentissage assisté par l'humain (Human-in-the-loop) pour guider efficacement la collecte et l'évaluation de la parole dans des scénarios multilingues hautement contraints.

5. https://semiolex.univ-lemans.fr/

Traitement du langage et humanités numériques

Avec des projets comme YLANG [6] nous proposons des modèles automatiques de reconnaissance de la parole qui sont capables de fournir des indices sur la prononciation en langue seconde (allemand, anglais) aux apprenants francophones. Les connaissances des difficultés de prononciation [33, 35] mais aussi de la prosodie [34] permettent d'ajuster les retours des systèmes aux difficultés de production réelles des apprenants et à l'acceptabilité des productions dans un cadre scolaire.

Depuis plusieurs années, LST a exploré les données d'archives médiatiques que ce soit pour des études historiques comme dans le projet ANTRACT [7] ou des études sociologiques comme dans le projet GEM évoqué plus haut. Dans les deux cas, les travaux ont mené à un large panel de modèles permettant le traitement automatique de données massives à des fins d'analyse. Cela permet d'étudier par exemple, l'évolution diachronique de la prononciation médiatique, ou bien le temps de parole des femmes et des hommes sur les médias français.

Nous exploitons également la question de l'interprétabilité des résultats ou des représentations des systèmes dans le cadre des humanités numériques. Cette question est en particulier centrale en scientométrie, où l'apprentissage non supervisé est mobilisé pour synthétiser de grandes collections de publications. Nous cherchons à travailler à l'évaluation de l'interprétabilité des résultats de ces algorithmes, et à la façon de les présenter de façon à favoriser la compréhension, mais avec fidélité. Ceci se retrouve également dans nos travaux en diachronie lexicale, lorsque nous interrogeons les dynamiques temporelles du langage.

Membres Impliqués

- Nicolas DUGUÉ (MCF)
- Anthony LARCHER (PR)
- Théo MARIOTTE (MCF)
- Sylvain MEIGNIER (PR)
- Aghilas SINI (MCF)
- Meysam SHAMSI (MCF)
- Marie TAHON (PR)
- Jane WOTTAWA (MCF)
- Kévin VYTHELINGUM (membre associé)

6. https://ylang.education/
7. https://lium.univ-lemans.fr/antract/

Références

[1] *TuniFra : A Tunisian Arabic Speech Corpus with Orthographic Transcriptions and French Translations*, Suzhou, China, 2025. Association for Computational Linguistics.

[2] Antonio Almudévar, Théo Mariotte, Alfonso Ortega, Marie Tahon, Luis Vicente, Antonio Miguel, and Eduardo Lleida. Predefined Prototypes for Intra-Class Separation and Disentanglement. In *Interspeech 2024*, pages 3809–3813, Kos, Greece, September 2024.

[3] Antonio Almudévar, Théo Mariotte, Alfonso Ortega, and Marie Tahon. Unsupervised multiple domain translation through controlled Disentanglement in variational autoencoder. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7010–7014, Seoul, France, April 2024. IEEE.

[4] Youness Dkhissi, Valentin Vielzeuf, Elys Allesiardo, and Anthony Larcher. SENS-ASR : Semantic Embedding injection in Neural-transducer for Streaming Automatic Speech Recognition. In *Language Resources and Evaluation Conference (LREC)*, Palma de Mallorca, Spain, May 2026.

[5] Thibault Gaudier, Marie Tahon, Anthony Larcher, and Yannick Estève. Automatic Voice Identification after Speech Resynthesis using PPG. In *Speaker and Language Recognition Workshop - Odyssey*, Québec (Canada), Canada, June 2024.

[6] Anthony Larcher, Yannick Estève, Mickael Rouvier, Natalia Tomashenko, Jarod Duret, Gaelle Laperriere, Santosh Kesijaru, Marek Sarvas, Renata Kohlova, Henry Li, Antoine Laurent, Thibault Gaudier, Valentin Pelloin, Thomas Thebaud, Emmanuelle Billard, Olivier Galibert, Swen Ribeiro, Harshita Diddee, Sameer Khurana, Laurent Besacier, Ioan Calapodescu, Themos Stafylakis, Tomas Pavli-

ceck, Cecile Macaire, Alejandro Ciuba, Peter Vickers, Luis Vicente, Victoria Mingote, and Pablo Gimeno. Multi-lingual Speech to Speech Translation for Under-Resourced Languages. Technical report, Le Mans Université, August 2022.

[7] Martin Lebourdais, Pablo Gimeno, Théo Mariotte, Marie Tahon, Alfonso Ortega, and Anthony Larcher. 3mas : a multitask, multilabel, multidataset semi-supervised audio segmentation model. In *Speaker and Language Recognition Workshop-Odyssey*, 2024.

[8] Martin Lebourdais, Marie Tahon, Antoine Laurent, and Sylvain Meignier. Automatic Speech Interruption Detection : Analysis, Corpus, and System. In *LREC-COLING)*, Torino, Italy, May 2024.

[9] Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau. AlloSat : A New Call Center French Corpus for Satisfaction and Frustration Analysis. In *LREC*, Marseille, France, May 2020.

[10] Théo Mariotte, Antonio Almudévar, Marie Tahon, and Alfonso Ortega. An explainable proxy model for multilabel audio segmentation. In *ICASSP*, pages 531–535. IEEE, 2024.

[11] Théo Mariotte, Anthony Larcher, Silvio Montrésor, and Jean-Hugh Thomas. Channel-Combination Algorithms for Robust Distant Voice Activity and Overlapped Speech Detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2024.

[12] Théo Mariotte, Martin Lebourdais, Antonio Almudévar, Marie Tahon, Alfonso Ortega, and Nicolas Dugué. Sparse autoencoders make audio foundation models more explainable. In *ICASSP*, pages 5256–5260. IEEE, 2026.

[13] Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Gabriel Hernández-Sierra, and Anthony Larcher. Lightweight Neural Networks for Multi-modal and Cross-modal Biometric Matching : Experimental Evaluation on Audio-Visual Person Verication. In *IX International Congress on Artificial Intelligence and Pattern Recognition IWAIPR 2025*, Varadero, Matanzas, Cuba, October 2025.

[14] Pauline Mas, Natacha Miniconi, Kevin Vythelingum, Meysam Shamsi, Aghilas Sini, Lu Zuo, Elias Okat, and Marie Tahon. Vo2lium : Voxygen and lium contribution for blizzard 2025. In *Blizzard Challenge*, 2025.

[15] Salima Mdhaffar, Valentin Pelloin, Antoine Caubrière, Gaëlle Laperriere, Sahar Ghannay, Bassam Jabaian, Nathalie Camelin, and Yannick Estève. Impact analysis of the use of speech and language models pretrained by self-supersivion for spoken language understanding. In *LREC*, pages 2949–2956, Marseille, France, June 2022.

[16] Sylvain Meignier and Teva Merlin. LIUM SPKDIARIZATION : AN OPEN SOURCE TOOLKIT FOR DIARIZATION. In *Proceedings CMU SPUD Workshop*, Dallas, United States, 2010.

[17] Natacha Miniconi, Meysam Shamsi, and Anthony Larcher. When the mos predictor asks for training annotation in cross lingual/domain adaptation. In *Interspeech*, 2025.

[18] Natacha Miniconi, Meysam Shamsi, Aghilas Sini, and Anthony Larcher. Using a deepfake classifier to rank speech synthesis quality. In *The Speech Synthesis Workshop (SSW)*, 2025.

[19] Mohammad Mohammadamini, Meysam Shamsi, and Marie Tahon. Central kurdish text-to-speech and its application in speech-to-text translation. In *LREC*, 2026.

[20] Mohammad Mohammadamini and Marie Tahon. Southern Kurdish speech recognition resources and benchmarking. In *LREC*, Palma, Mallorca, Spain, May 2026.

[21] Mohammad Mohammadamini, Marie Tahon, Aghilas Sini, and Antoine Laurent. Apprentissage de modèles frugaux pour les langues peu dotées à partir de larges modèles d'ASR. In *JEP*, Montpelier, France, June 2026. AFCP.

[22] Hubert Nourtel, Pierre Champion, Denis Jouvet, Anthony Larcher, and Marie Tahon. Evaluation of Speaker Anonymization on Emotional Speech. In *SPSC 2021 - 1st ISCA*

Symposium on Security and Privacy in Speech Communication, Virtual, Germany, November 2021.

[23] Valentin Pelloin, Nathalie Camelin, Antoine Laurent, Renato de Mori, Antoine Caubrière, Yannick Estève, and Sylvain Meignier. End2End Acoustic to Semantic Transduction. In *ICASSP*, Toronto, ON, Canada, June 2021.

[24] David Perera, François Derrida, Théo Mariotte, Gaël Richard, and Slim Essid. Multiple choice learning for efficient speech separation with many speakers. In *ICASSP*, pages 1–5. IEEE, 2025.

[25] Félix Saget, Nicolas Dugué, Marie Tahon, and Anthony Larcher. Évaluation perceptive de l'interprétabilité dimensionnelle dans les représentations parcimonieuses du locuteur. In *JEP2026*, Montpellier, France, June 2026.

[26] Félix Saget, Thibault Gaudier, Meysam Shamsi, and Marie Tahon. Lium-tts entry for blizzard 2023. In *Blizzard Challenge Workshop*, pages 28–33. ISCA, 2023.

[27] Félix Saget, Nicolas Dugué, Marie Tahon, and Anthony Larcher. Functionnally-grounded evaluation of dimensional interpretability in sparse speaker representations. In *Speaker Odyssey*, Lisboa, Portugal, 2026. ISCA.

[28] Meysam Shamsi, Anthony Larcher, Loïc Barrault, Sylvain Meignier, Yevheni Prokopalo, Marie Tahon, Ambuj Mehrish, Simon Petitrenaud, Olivier Galibert, Samuel Gaist, André Anjos, Sébastien Marcel, and Marta Costa-Jussà. Towards Lifelong Human Assisted Speaker Diarization. *Computer Speech and Language*, 77 :101437, 2023.

[29] Marie Tahon, Manon Macary, and Yannick Estève. Continuous emotion prediction from audio signal with acoustic and linguistic representations. In *16ème Congrès Français d'Acoustique, CFA2022*, Marseille, France, April 2022. Société Française d'Acoustique and Laboratoire de Mécanique et d'Acoustique.

[30] Hoan My Tran, Damien Lolive, David Guennec, Aghilas Sini, Arnaud Delhay, and Pierre-François Marteau. Leveraging SSL Speech Features and Mamba for Enhanced DeepFake Detection. In *Interspeech*, pages 5323–5327, Rotterdam (NL), Netherlands, August 2025.

[31] Hoan My Tran, Aghilas Sini, David Guennec, Arnaud Delhay, Damien Lolive, and Pierre-François Marteau. Kan we make models simpler for audio deepfake detection with kolmogorov–arnold networks ? In *ICASSP*, pages 18057–18061, 2026.

[32] Rémi Uro, Marie Tahon, Jane Wottawa, David Doukhan, Albert Rilliard, and Antoine Laurent. Annotation of transition-relevance places and interruptions for the description of turn-taking in conversations in French media content. In *LREC-COLING*, Torino, Italia, May 2024.

[33] Jane Wottawa and Martine Adda-Decker. Quand les voyelles longues et brèves ne tiennent pas en place : la qualité vocalique en allemand L2. In *JEP*, pages 64–71, Aix-en-Provence, France, June 2018. ISCA.

[34] Jane Wottawa and Martine Adda-Decker. Lecture L2 : modulations prosodiques aux frontières syntaxiques majeures de textes allemands. In *JEP*, Montpellier, France, June 2026. ISCA.

[35] Jane Wottawa, Maelle Amand, and Takeki Kamiyama. To [h] or not to [h] ? L2 production of /h/ in semi-spontaneous speech by French leaners of English and of German. In *L2APR : L2 ACCENT AND PRONUNCIATION RESEARCH : ACQUISITION, TEACHING, ATTITUDES*, Venise, Italy, November 2023.

AfIA
Association française
pour l'Intelligence Artificielle

LLF : Laboratoire de Linguistique Formelle

LLF UMR 7110
CNRS et Université Paris Cité
http://www.llf.cnrs.fr/fr/themes_de_recherche

Thimothée BERNARD
Marie CANDITO
Benoît CRABBÉ
Guillaume WISNIEWSKI
prenom.nom@u-pariscite.fr

Présentation générale

L'axe "linguistique computationnelle" du laboratoire LLF (Université Paris Cité et CNRS) est porté par quatre chercheur·euse·s permanents, aidés par actuellement 5 doctorant·e·s et deux post-doctorant·e·s. Ce groupe se consacre à l'étude du traitement automatique des langues et à la linguistique computationnelle. Il bénéficie d'une forte intégration dans le paysage scientifique local, avec des collaborations privilégiées avec les laboratoires d'informatique de l'université : l'IRIF avec l'ERC Malinca et le LIPADE avec le projet ANR Grail. Par l'intermédiaire de l'InidEx EFL[8] il y a également des interactions avec des équipes de linguistique de l'Université Paris Cité, comme ALTAE, de la Sorbonne Nouvelle et du laboratoire d'informatique de Paris Nord (LIPN).

Le groupe est activement impliqué dans plusieurs projets de recherche, tant nationaux qu'internationaux. Parmi ces initiatives, on trouve des projets tels que l'ANR Compo, qui explore les modèles de généralisation et compositionnalité en apprentissage artificiel et en linguistique, l'ANR DeepTypo, axé sur l'extraction d'informations liées à la typologie des langues à partir des représentations neuronales, l'ANR SELEXINI, visant à l'induction automatique de sens et de cadres sémantiques, l'ANR Pantagruel, qui développe des modèles de langue multimodaux (texte, audio et pictogrammes).

Principaux thèmes de recherche

Le groupe explore différents aspects du TAL, avec un accent particulier sur l'apport d'un regard linguistique dans le développement et l'évaluation des grands modèles de langue (nous avons notamment participé au développement de FlauBERT et de Pantagruel), la traduction automatique [11].

Historiquement, le groupe a eu une activité importante dans le domaine de l'analyse syntaxique et sémantique automatique (par exemple [3, 7]). Cette activité a évolué vers des thématiques d'IA neuro-symbolique liées à la volonté d'apporter des garanties de certification quant aux activités liées au langage et au raisonnement en IA. Il s'agit de prédire des structures discursives, temporelles ou argumentatives à partir de textes dans le but de produire des représentations de données pour lesquelles il est possible de certifier le raisonnement ou les inférences qui en sont tirées (comme la preuve mathématique ou le raisonnement non monotone mené dans certaines institutions sociétales).

Une part importante du travail du groupe est dédiée à l'analyse des représentations neuronales, afin de mieux comprendre comment les LLM manipulent la langue. Par exemple, des recherches ont été menées sur des phénomènes linguistiques tels que la négation [9] et l'accord du participe passé [10], afin de déceler les mécanismes internes qui permettent à ces modèles de gérer des règles grammaticales complexes. Récemment, cette thématique s'est élargie pour inclure l'analyse des représentations neuronales dans le domaine de l'audio [17], ainsi que l'analyse de l'organisation des informations lexicales, via l'étude de la tâche d'associations de mots [16].

Plus généralement, le succès pratique des modèles neuronaux de traitement du langage soulève de nombreuses questions en linguistique théorique et philosophie du langage. À l'inverse, ces modèles peuvent être mis à profit pour développer des outils utiles à l'étude de la syntaxe, la sémantique, de l'ac-

8. https://www.labex-efl.fr/

quisition du langage... Ces questions sont explorées par certains membres du groupe [6], en particulier dans le séminaire ILFC (Interactions entre linguistique formelle et computationnelle), ainsi qu'une série d'ateliers "BriGap" (Bridges and Gaps between Formal and Computational Linguistics).

Le groupe étudie également les jeux d'émergence du langage, dont le but à long terme de comprendre dans quelles conditions les signaux échangés par les agents acquièrent certaines propriétés du langage humain — notamment, une structure syntaxique et une sémantique compositionnelle [4, 5].

Outre l'analyse de modèles neuronaux, certains travaux de l'équipe visent à l'amélioration de ceux-ci, par exemple concernant l'amélioration de représentations neuronales de sens [12, 18], en vue notamment d'améliorer l'induction automatique de sens [13, 19].

Nous menons en outre des travaux sur la notion de généralisation en apprentissage profond. Ce travail comporte deux aspects. D'une part un aspect théorique dédié à l'étude de la notion de causalité dans le contexte de l'apprentissage en partenariat avec l'entreprise OnePoint. Dans un premier temps l'idée est d'identifier des méthodes destinées à privilégier les causes réelles plutôt que des corrélats fallacieux lors de l'apprentissage de modèles ce qui peut mener à terme à des méthodes d'apprentissage de modèles moins sensibles aux biais (comme biais de race ou de genre). Nous menons d'autre part un travail en lien avec la simulation de capacités de généralisation chez les humains. Cela implique de créer et d'évaluer des modèles qui simulent les limitations de mémoire [14] typiques dans le traitement du langage chez l'humain et également d'évaluer les modèles existants sur leur capacité à apprendre des généralisations subtiles qui dépassent les règles de grammaire classiques, comme des généralisations en lien avec la notion de *construction* en linguistique [15].

Un autre axe de développement de l'équipe depuis plusieurs années est la documentation linguistique computationnelle, qui vise à utiliser les outils de TAL pour aider les linguistes dans leur travail de documentation et de préservation des langues. Ce domaine inclut le développement de systèmes de transcription adaptés aux contextes à faibles ressources [1]. Nous travaillons également sur des outils permettant d'identifier les caractéristiques spécifiques des langues, ainsi que des solutions pour comparer des langues entre elles, afin de mieux comprendre leurs structures [8]. Ce thème vise aussi des langues très dotées comme le français, pour lequel nous avons par exemple travaillé à typer sémantiquement des entrées lexicales [2].

Membres permanents

- Timothée BERNARD (MCF)
- Marie CANDITO (MCF, HDR)
- Benoît CRABBÉ (PR)
- Guillaume WISNIEWSKI (MCF)

Références

[1] Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques, and Nathan Hill. User-friendly automatic transcription of low-resource languages : Plugging ESPnet into elpis. In Antti Arppe, Jeff Good, Atticus Harrigan, Mans Hulden, Jordan Lachler, Sarah Moeller, Alexis Palmer, Miikka Silfverberg, and Lane Schwartz, editors, *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 51–62, Online, March 2021. Association for Computational Linguistics.

[2] Nicolas Angleraud, Lucie Barque, and Marie Candito. Annotating the French Wiktionary with supersenses for large scale lexical analysis : a use case to assess form-meaning relationships within the nominal lexicon. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5321–5332, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[3] Timothée Bernard. Multiple tasks integration : Tagging, syntactic and semantic parsing

as a single task. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 783–794, Online, April 2021. Association for Computational Linguistics.

[4] Timothée Bernard and Timothee Mickus. So many design choices : Improving and interpreting neural agent communication in signaling games. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics : ACL 2023*, pages 8399–8413, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Timothée Bernard, Timothee Mickus, and Hiroya Takamura. The emergence of high-level semantics in a signaling game. In Danushka Bollegala and Vered Shwartz, editors, *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 200–211, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[6] Timothée Bernard and Grégoire Winterstein. Introduction to the special section on the interaction between formal and computational linguistics. *Journal of Language Modelling*, 10(1) :39–47, Oct. 2022.

[7] Marie Candito. Auxiliary tasks to boost biaffine semantic dependency parsing. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics : ACL 2022*, pages 2422–2429, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] Maxime Fily, Guillaume Wisniewski, Severine Guillaume, Gilles Adda, and Alexis Michaud. Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics : EACL 2024*, pages 2332–2341, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[9] David Kletz, Pascal Amsili, and Marie Candito. Polarity inversion operators in PLM. In Gemma Boleda and Michael Roth, editors, *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 312–322, Vienna, Austria, July 2025. Association for Computational Linguistics.

[10] Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. Assessing the capacity of transformer to abstract syntactic representations : A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, 11 :18–33, 2023.

[11] Joachim Minder, Guillaume Wisniewski, and Natalie Kübler. Testing LLMs' capabilities in annotating translations based on an error typology designed for LSP translation : First experiments with ChatGPT. In Pierrette Bouillon, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX : Volume 1*, pages 190–203, Geneva, Switzerland, June 2025. European Association for Machine Translation.

[12] Anna Mosolova, Marie Candito, and Carlos Ramisch. Injecting Wiktionary to improve token-level contextual representations using contrastive learning. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 34–41, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[13] Anna Mosolova, Marie Candito, and Carlos Ramisch. In the LLM era, word sense induction remains unsolved. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics : ACL 2025*, pages 17161–17178, Vienna, Austria, July 2025. Association for Computational Linguistics.

[14] Nina Nusbaumer, Corentin Bel, Iria de Dios-Flores, G Wisniewski, and B Crabbé. Dependency length, syntactic complexity and memory : A reading time benchmark for sentence processing modeling. In *Computational Psycholinguistics Meeting 2025*, 2025.

[15] Nina Nusbaumer, G Wisniewski, and B Crabbé. Évaluer la capacité des transformeurs à distinguer les significations compositionnelles et idiomatiques d'une même expression. In *Proc TALN 2025*, 2025.

[16] Maria A. Rodriguez, Marie Candito, and Richard Huyghe. Modeling the human lexicon under temperature variations : Linguistic factors, diversity and typicality in llm word associations. In *To appear in the LREC 2026 proceedings*, May 2026.

[17] Eli Stafford, Aimée Lahaussois, and Guillaume Wisniewski. What do neural speech models know about phonology ? evidence from structured phoneme confusions. In *Findings of the Association for Computational Linguistics : ACL 2026*, 2026.

[18] Deniz Ekin Yavas, Timothée Bernard, Benoit Crabbé, and Laura Kallmeyer. On the relation between fine-tuning, topological properties, and task performance in sense-enhanced embeddings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 23610–23625, Vienna, Austria, July 2025. Association for Computational Linguistics.

[19] Deniz Ekin Yavas, Timothée Bernard, Laura Kallmeyer, and Benoît Crabbé. Improving word sense induction through adversarial forgetting of morphosyntactic information. In Danushka Bollegala and Vered Shwartz, editors, *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 238–251, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

AfIA
Association française
pour l'Intelligence Artificielle

LORIA/Sémagramme : Analyse sémantique du langage naturel

LORIA
INRIA et Université de Lorraine
https://team.inria.fr/semagramme/research/

Maxime AMBLARD
Responsable d'équipe
maxime.amblard@loria.fr

Vincent P. MARTIN
Coordinateur du bulletin
vincent.p.martin@inria.fr

Présentation générale

L'équipe Sémagramme du LORIA a pour objectif de mettre au jour une compréhension fine du fonctionnement du langage dans un objectif de Traitement Automatique des Langues (TAL), en développant des méthodes pour l'analyse de modèles, en concevant des algorithmes pour identifier les structures linguistiques, en créant des outils et en évaluant les capacités sémantiques, biais et potentielles applications, tout en incorporant des réflexions éthiques et critiques sur leur utilisation. A travers cette approche, l'équipe désire contribuer au développement et à l'évaluation de modèles qui sont transparents, interprétables, et attentifs aux enjeux sociétaux et cognitifs du langage. Les outils développés ont pour objectif de soutenir à la fois le développement de modèles exacts et justes et leur évaluation qualitative.

Principaux thèmes de recherche

Le travail de l'équipe est organisé autour d'un objectif global : développer des algorithmes, méthodes et outils pour la compréhension des capacités, limites et l'interprétabilité des modèles de TAL, avec un focus sur le traitement de la sémantique et du discours. Cet objectif est adressé à travers quatre axes thématiques.

Axe 1 - Ressources Linguistiques. L'équipe porte une attention particulière aux données linguistiques riches, représentatives de la diversité des langues et de leurs usages. Il ne s'agit pas seulement d'élargir leur couverture, mais aussi de repenser la construction, l'annotation et l'exploitation des ressources linguistiques grâce à des approches symboliques, qui offrent pour cela des solutions sobres et interprétables, adaptées à l'analyse fine et à la manipulation explicite des représentations. Le logiciel Grew, consacré à la recherche et à la réécriture de graphes, illustre ce potentiel dans des projets multilingues tels qu'*Universal Dependencies* [13].

De plus, nous concevons des ressources capables d'étayer des hypothèses linguistiques complexes. Leur production suppose de transformer des annotations structurées, d'aligner plusieurs niveaux d'analyse (syntaxe, sémantique, pragmatique) et de construire des ressources multilingues comparables. Nos choix méthodologiques partent des besoins d'applications concrètes afin de construire des chaînes de traitement robustes et transparentes. Les annotateurs jouent ici un rôle essentiel : leur expertise fonde les interprétations sur lesquelles reposent les traitements automatiques. Notre méthode combine des guides d'annotation adaptés aux données, des cycles de validation et de révision, ainsi que des outils dédiés à la transformation et à l'exploitation des structures [14, 4]. Elle fournit ainsi un cadre opérationnel pour produire des données annotées de qualité, dans de nombreuses langues et pour différents phénomènes linguistiques.

Axe 2 - Évaluation et analyse éthique. Un enjeu majeur de la recherche en traitement automatique des langues consiste à éviter de prêter aux grands modèles de langue des capacités qui dépassent leur conception et leurs performances réelles, notamment lorsqu'ils sont employés comme sources d'information ou dans des contextes sensibles, tels que la médecine, sans garantie d'exactitude ni évaluation adaptée aux usages [3]. Similairement, l'apparition et la propagation des biais issus des corpus, des annotations, des architectures et des procédures d'adaptation, doivent être identifiées et quantifiées,

spécifiquement selon les langues, les domaines et les applications de ces modèles [12]. Les métriques classiques et les *benchmarks* ne donnent en effet qu'une vision partielle des modèles, car ils rendent mal compte de leur robustesse, de leur interprétabilité, de leur diversité linguistique et de leur compréhension du contexte [17]. Nous cherchons donc à élaborer des méthodes d'évaluation systématiques combinant mesures quantitatives, analyses diagnostiques, tests de résistance et approches qualitatives, tout en développant des ressources pour évaluer ou adapter les modèles [11, 5] et des recommandations intégrant les enjeux éthiques [1]. Cette démarche doit aussi prendre en compte les coûts énergétiques et les compromis entre performance, efficacité et impact environnemental, afin de soutenir la conception de systèmes plus équitables et contrôlables, ainsi que leur audit et leur régulation selon des exigences de transparence, de robustesse et de responsabilité.

Axe 3 - Représentations sémantiques. La représentation sémantique demeure un enjeu central du traitement automatique des langues pour concevoir des systèmes robustes, interprétables et capables de traiter explicitement le sens. Notre objectif est de développer des représentations à la fois calculables, linguistiquement fondées et compatibles avec les méthodes modernes d'apprentissage automatique, en nous appuyant sur la linguistique formelle, la logique et l'informatique théorique. Nos recherches portent notamment sur la compositionnalité, encore difficile à maîtriser dans les grands modèles de langue, et mobilisent les grammaires catégorielles abstraites, l'isomorphisme de Curry–Howard et la théorie des catégories afin d'articuler analyses syntaxiques formelles et espaces sémantiques distributionnels [7, 23]. Elles étudient également la signification des relations observées dans les espaces vectoriels issus des plongements lexicaux et leur mise en correspondance avec des relations conceptuelles définies par la linguistique ou la logique [16]. Dans cette perspective, le formalisme Yarn, développé par Maxime Amblard et Bruno Guillaume, associe la souplesse des représentations en graphes à la rigueur des approches logiques afin de prendre en charge l'annotation, le raisonnement, la vérification et la transformation contrôlée de phénomènes sémantiques variés [22]. La réécriture de graphes, notamment au moyen de grew, permet alors de relier systématiquement différents niveaux de représentation, de la syntaxe aux formes logiques et discursives, tandis que des questions comme la portée de la négation et de la quantification restent des défis majeurs. Enfin, nous développons une chaîne de traitement hybride combinant apprentissage neuronal et contraintes symboliques pour produire automatiquement des représentations Yarn à partir de données parfois bruitées ou incomplètes, les évaluer au moyen de métriques adaptées, puis les transformer en formules logiques destinées au raisonnement [18, 8, 9, 10] ; les travaux futurs viseront à affiner ces représentations et les modèles d'analyse, à systématiser leur évaluation et à renforcer les modules de transformation, de validation et d'inférence.

Axe 4 - Cohésion du discours et cohérence. Le quatrième axe, au cœur du projet, étudie la capacité des outils de traitement automatique des langues à intégrer le contexte au sens large, en particulier dans le discours et le dialogue, où le sens se construit progressivement grâce à la continuité thématique, à la progression logique, aux dynamiques interactionnelles et aux intentions communicatives. Il s'agit ainsi de dépasser une analyse purement sémantique pour comprendre comment les modèles représentent et exploitent les dimensions contextuelles et pragmatiques qui fondent la cohérence. Le projet ERC MALINCA illustre cette démarche par l'étude du discours mathématique et par la création de ressources annotées et de modèles capables d'identifier automatiquement ses structures argumentatives, leurs relations et leurs ruptures. L'axe examine également les marqueurs linguistiques de cohérence — connecteurs, anaphores, temps verbaux, reformulations, ruptures et reprises [6] — en s'appuyant notamment sur la RST, la SDRT et des approches empiriques sur corpus. Enfin, l'analyse du dialogue prend en compte les stratégies interactionnelles, les intentions partagées ou implicites ainsi que l'influence des états mentaux, du regard et de l'expression vocale [25, 2], avec des applications possibles dans des domaines tels que la santé men-

tale. L'objectif général est de déterminer comment la cohérence se construit au-delà de la phrase, comment intégrer ces informations aux modèles computationnels et comment évaluer ces mécanismes de manière fiable et interprétable.

Applications

Ces quatre axes de recherche s'illustrent à travers 3 champs d'applications.

Discours mathématique. Dans le cadre du projet ERC Synergy MALINCA (MAthematicae LINgua franCA), Sémagramme étudie le traitement automatique du discours mathématique, qui associe des notations rigoureuses à une langue naturelle nuancée. L'objectif est de concevoir des représentations intermédiaires capables de restituer à la fois la structure logique et la richesse linguistique des définitions, théorèmes et démonstrations, afin de rapprocher le raisonnement humain des assistants de preuve tels que Rocq (anciennement Coq). Ces travaux pourraient faciliter la vérification automatique des preuves, offrir une aide personnalisée à leur rédaction et rendre les outils de formalisation plus accessibles aux chercheurs, aux enseignants et aux étudiants.

Régulation de l'IA. L'équipe développe des méthodes d'évaluation des modèles de langue et des systèmes d'intelligence artificielle afin d'en mesurer la fiabilité, l'équité, la transparence et la responsabilité, notamment au regard du règlement européen sur l'intelligence artificielle. Des protocoles combinant benchmarks diagnostiques, analyse des biais, métriques, cadres d'annotation et expérimentations contrôlées permettent de mieux caractériser les capacités et les limites de ces systèmes dans des situations réalistes. Ces ressources soutiennent à la fois la recherche et les démarches d'audit, de certification, de conformité réglementaire et d'évaluation des risques menées par les institutions publiques, les autorités de régulation et les acteurs industriels.

Santé. Dans le domaine de la santé, l'équipe analyse les marqueurs linguistiques, discursifs et vocaux susceptibles de contribuer à la détection et au suivi de pathologies, en particulier dans le champ de la santé mentale. L'étude de réseaux de symptômes [21] et de biomarqueurs linguistiques ou acoustiques [20], à partir notamment de plus de 200 entretiens cliniques recueillis avec plusieurs hôpitaux français et algériens, vise à repérer précocement des signes de troubles tels que la dépression ou la schizophrénie [15, 19]. En complément de l'expertise clinique, ces méthodes non invasives pourraient améliorer le suivi des patients et la personnalisation des interventions ; elles sont également explorées avec des services d'urgence pour détecter rapidement, dans la voix, des indices de stress, de fatigue ou de défaillance cardiaque [24].

Membres Permanents

- Maxime AMBLARD (PR)
- Julien BEZANÇON (MCF)
- Mathieu CONSTANT (Membre associé, PR)
- Karën FORT (PR)
- Mathilde DARGNAT (Membre associé, PR)
- Philippe DE GROOTE (DR INRIA)
- Bruno GUILLAUME (CR INRIA)
- Jacques JAYEZ (Emerite, PR)
- Vincent P. MARTIN (CR INRIA)
- Michel MUSIOL (Membre associé, PR)
- Sylvain POGODALLA (CR INRIA)

Références

[1] Marc M Anderson and Karën Fort. Evaluating the acceptability of ethical recommendations in industry 4.0 : an ethics by design approach. *AI & Society : Knowledge, Culture and Communication*, January 2024.

[2] Vincent-Thomas Barrouillet, Maxime Amblard, Sadeq Haouzir, Thibault Delage, and Michel Musiol. L'hypothèse du continuum bipolarité-schizophrénie au risque psycholinguistique des ruptures discursives. *Annales Médico-Psychologiques, Revue Psychiatrique*, January 2025.

[3] Clémentine Bleuze, Karën Fort, Vincent P Martin, and Aurélie Névéol. Large Language Models for Mental Health Prediction : Scoping Review of Bias and Clinical Utility Documentation in 2019-2024. *JMIR AI*, 5 :e88082–e88082, August 2026.

[4] Harry Bunt, Maxime Amblard, Johan Bos, Karën Fort, Philippe de Groote, Bruno Guillaume, Chuyuan Li, Pierre Ludmann, Michel Musiol, Siyana Pavlova, Guy Perrier, and Sylvain Pogodalla. Quantification Annotation in ISO 24617-12, Second Draft. In Nicoletta Calzolari, Frédéric Béchet, Hélène Mazo, Jan Odijk, Stelios Piperidis, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*, pages 3407–3416, Marseille, France, June 2022. European Language Resources Association.

[5] Hee-Soo Choi, Priyansh Trivedi, Mathieu Constant, Karën Fort, and Bruno Guillaume. Au-delà de la performance des modèles : la prédiction de liens peut-elle enrichir des graphes lexico-sémantiques du français ? In Mathieu Balaguer, Nihed Bendahman, Lydia-Mai Ho-Dac, Julie Mauclair, Jose G. Moreno, and Julien Pinquier, editors, *Actes de JEP-TALN-RECITAL 2024. 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 36–49, Toulouse, France, July 2024. ATALA & AFPC.

[6] Mathilde Dargnat. Mais enfin : construction et association. *Langages*, 1(225) :49–63.

[7] Philippe de Groote. Towards Abstract Categorial Grammars. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 148–155, July 2001.

[8] Rémi de Vergnette and Maxime Amblard. Semantic Parsing for Evaluating Large Language Models : Separating Linguistic Abilities with YARN. In *LREC 2026 - Language Resources and Evaluation Conference*, Palma De Majorque, Spain, May 2026.

[9] Rémi de Vergnette, Maxime Amblard, and Bruno Guillaume. Evaluation framework for layered meaning representation. In Kenneth Lai and Shira Wein, editors, *Proceedings of the Sixth International Workshop on Designing Meaning Representations*, pages 38–48, Prague, Czechia, aug 2025. Association for Computational Lingustics.

[10] Rémi de Vergnette, Vincent Tourneur, and Maxime Amblard. Extracting First Order Logic formulas from graphical semantic representations. In *DMR 2026 - 7th International Workshop on Designing Meaning Representations*, Palma de majorque, Spain, May 2026.

[11] Fanny Ducel, Jeffrey André, Aurélie Névéol, and Karën Fort. Introducing MascuLead : the First Gender Bias Leaderboard. In *Actes de l'atelier Ethic and Alignment of (Large) Language Models 2025 (EALM)*, pages 12–19, Marseille, France, June 2025.

[12] Fanny Ducel, Aurélie Névéol, and Karën Fort. "You'll be a nurse, my son !" Automatically Assessing Gender Biases in Autoregressive Language Models in French and Italian. *Language Resources and Evaluation*, October 2024.

[13] Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, and Yixuan Li. Joint annotation of morphology and syntax in dependency treebanks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9568–9577, Torino, Italia, May 2024. ELRA and ICCL.

[14] Nicolas Hiebel, Bertrand Remy, Bruno Guillaume, Olivier Ferret, Aurélie Névéol, and Karën Fort. Hostomytho : A GWAP for Synthetic Clinical Texts Evaluation and Annotation. In *10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024*, pages 14–20, Turin, Italy, May 2024.

[15] Alexandra König, Philipp Müller, Johannes Tröger, Hali Lindsay, Jan Alexandersson, Jonas Hinze, Matthias Riemenschneider, Danilo Postin, Eric Ettore, Amandine Lecomte, Michel Musiol, Maxime Amblard, François Bremond, Michal Balazia, and Rene Hurlemann. Multimodal phenotyping of psychiatric disorders from social interaction : Protocol of a cli-

nical multicenter prospective study. *Personalized Medicine in Psychiatry*, 33-34 :100094, May 2022.

[16] Timothee Mickus, Denis Paperno, and Mathieu Constant. How to dissect a muppet : The structure of transformer embedding spaces. *Transactions of the Association for Computational Linguistics*, 10 :981–996, 09 2022.

[17] Alice Millour, Yoann Dupont, Karën Fort, and Liam Duignan. Unveiling Strengths and Weaknesses of NLP Systems Based on a Rich Evaluation Corpus : the Case of NER in French. In *Proceedings of LREC-COLING 2024*, Turin, Italy, May 2024.

[18] Ivaylo Mitov, Tadzhat Marharian, Zsofia F. Hauk, Samba Fall, Maxime Amblard, and Bruno Guillaume. AMR Parsing beyond English : An Experiment on Bulgarian, French, Hungarian and Ukrainian. In *LREC 2026 - Language Resources and Evaluation Conference*, Palma De Majorque, Spain, May 2026.

[19] Michel Musiol, Arthur Trognon, and Maxime Amblard. L'analyse de l'interaction verbale " patient " - " thérapeute " par la modélisation formelle : perspectives diagnostiques et informatisation. In Yann Auxéméry and Jasmina Mallet, editors, *Psychiatrie et Psychologie du Futur*, pages 83–106. Ellipses, June 2025.

[20] Vincent P. Martin, Charles Brazier, Maxime Amblard, Michel Musiol, and Jean-Luc Rouas. Network of acoustic characteristics for the automatic detection of suicide risk from speech. Contribution to the 2025 SpeechWellness challenge by the Semawave team. In Odette Scharenborg, Catharine Oertel, and Khiet Truong, editors, *Interspeech 2025 : proceeding*, pages 424–428, Rotterdam (NL), Netherlands, August 2025. ISCA.

[21] Vincent P. Martin, Christophe Gauld, and Sébastien Bailly. Gentle Introduction to Network Analysis for Clinical Research : a model centered on relations. *Chest*, 168(3) :574–577, September 2025.

[22] Siyana Pavlova, Maxime Amblard, and Bruno Guillaume. YARN is All You Knit : Encoding Multiple Semantic Phenomena with Layers. In *Proceedings of The Fifth International Workshop in Designing Meaning Representation*, Turin, Italy, May 2024.

[23] Sylvain Pogodalla. A syntax-semantics interface for Tree-Adjoining Grammars through Abstract Categorial Grammars. *Journal of Language Modelling*, 5(3) :527–605, 2017.

[24] Elio Stasica, Clément Joly, Amandine Lecomte, Vincent P. Martin, Romain Serizel, Emmanuel Vincent, and Tahar Chouihed. Revisiting Emotion-Based Triage : Evidence from French Emergency Call Data. In *Interspeech*, Sidney, Australia, September 2026.

[25] Arthur Trognon, Camille Humeau, Loann Mahdar-Recorbet, Frédéric Verhaegen, and Michel Musiol. A physical framework to harmonize human interaction analysis across disciplines. *Current Psychology*, 44(5) :3519–3531, 2025.

AfIA
Association française
pour l'Intelligence Artificielle

LS2N/TALN : Traitement Automatique du Langage Naturel

LS2N UMR 6004
CNRS et Nantes Université
https://taln.ls2n.fr/

Florian BOUDIN
Richard DUFOUR
prenom.nom@ls2n.fr

Présentation générale

L'équipe TALN est une équipe de recherche du LS2N (UMR 6004, CNRS et Nantes Université) dont les travaux s'inscrivent dans le champ du traitement automatique des langues (TAL). Ses recherches couvrent un large spectre de la discipline, de la modélisation du langage à l'IA générative, en passant par l'extraction d'information, la collecte et la structuration de données, la construction de ressources et de benchmarks, ainsi que l'évaluation des systèmes. Les travaux de l'équipe se caractérisent par une articulation forte entre avancées méthodologiques, production de ressources et développement d'applications. Ils portent à la fois sur la conception de modèles et de méthodes pour analyser, représenter et générer des contenus langagiers, et sur la mise à disposition de données, corpus et protocoles d'évaluation utiles à la communauté scientifique. L'équipe développe ses recherches dans plusieurs domaines applicatifs à fort impact sociétal, en particulier la santé et le biomédical, la recherche scientifique, l'éducation et le droit. Dans ces différents contextes, elle s'intéresse à des problématiques où les usages, la fiabilité des systèmes et l'adaptation aux spécificités des données jouent un rôle central.

Principaux thèmes de recherche

TAL pour la santé et le biomédical

L'équipe développe des méthodes de TAL appliquées aux données de santé, avec un accent particulier sur les approches de science ouverte (modèles, données, outils). Ces travaux s'inscrivent dans un contexte fortement contraint, caractérisé par l'hétérogénéité et le bruit des données, ainsi que des exigences élevées de confidentialité, de robustesse et d'interprétabilité.

Nos travaux portent d'abord sur la conception de modèles de langue spécialisés pour le domaine biomédical, avec un double objectif : fournir à la communauté des modèles et outils ouverts, et maîtriser l'ensemble de la chaîne de développement, des données à l'entraînement des modèles. L'équipe a entraîné et diffusé librement plusieurs modèles préentraînés en français adaptés aux données de santé, notamment *DrBERT* [15], entraîné sur des corpus biomédicaux ouverts, ainsi que des modèles génératifs tels que *BioMistral*, déclinés en versions avec entraînement continu [17] et avec instruction [5], conçus pour des usages avancés de compréhension et de génération de textes médicaux, avec une adaptation sur des données en anglais et en français. Ces travaux visent à mieux prendre en compte les spécificités lexicales, terminologiques et discursives du domaine, tout en étudiant les verrous liés à l'adaptation des grands modèles de langue à des données spécialisées (e.g. impact de l'utilisation de données publiques *vs.* privées [15], de la tokenisation [14], transfert entre langues [17], biais [19]). Les modèles ouverts BioMistral et DrBERT ont eu un impact important dans la communauté, avec un cumul de plus de 1,4 millions de téléchargements et plus de 600 citations entre 2023 et 2025.

L'équipe accorde également une place importante à l'évaluation. Ces travaux ont conduit à la conception de benchmarks dédiés, tels que *DrBenchmark* [16], permettant d'évaluer de manière systématique les capacités de compréhension des modèles sur de très nombreuses tâches biomédicales en français (e.g. classification, reconnaissance d'entités, question-réponse...). Un enjeu central est la définition de protocoles d'évaluation permettant l'analyse de la robustesse des modèles, de leur généralisabilité et de leur fiabilité dans des contextes biomédicaux variés. Nous étudions également des approches d'évaluation automatique fondées sur des modèles de type *LLM-as-a-judge* [4].

Enfin, l'équipe développe des approches pour l'exploitation des données textuelles en contexte cli-

nique et en santé publique. Ces travaux, menés notamment dans le cadre de projets structurants tels que le projet *ANR MALADES* et le projet *BPI Partages*, visent à développer des solutions concrètes et souveraines pour le TAL médical en contexte contraint. Ils couvrent des problématiques telles que l'extraction d'informations cliniques [1], l'analyse de parcours de soins ou encore l'intégration de données issues des patients [2]. Ces recherches donnent lieu à des applications en collaboration avec des acteurs hospitaliers, en particulier le CHU de Nantes, pour l'analyse de comptes rendus cliniques et l'aide à l'exploitation de données patients à des fins de recherche clinique et d'amélioration des pratiques de soins. Elles soulèvent des verrous scientifiques liés à la qualité et à la représentativité des données, à l'intégration de sources multimodales [18], ainsi qu'à la mise en place de méthodes fiables pour des usages en appui à la décision médicale [13].

TAL pour la recherche scientifique

L'équipe conçoit des méthodes pour assister les chercheurs dans leur interaction avec la littérature scientifique, de son exploration à son analyse, sa rédaction et la vérification des résultats. Ses travaux portent d'abord sur l'accès à la littérature, avec des recherches sur l'enrichissement de l'indexation par génération automatique de mots-clés [7], la conception de moteurs de recherche et de systèmes de recommandation adaptés aux usages scientifiques [6].

Elle développe également des travaux sur l'aide à l'écriture scientifique. Ceux-ci portent notamment sur l'identification de formulations récurrentes dans les articles [11], utiles pour expliciter les contributions et structurer le discours, ainsi que sur l'évaluation de systèmes de révision automatique fondés sur les grands modèles de langue [12], afin d'améliorer la clarté des textes sans altérer l'intention de l'auteur ni le contenu scientifique.

Enfin, l'équipe s'intéresse à la vérification des affirmations scientifiques, dans une perspective de rigueur et de fiabilité des publications. Ces travaux visent à développer des méthodes capables de relier les affirmations aux éléments qui les étayent dans les articles, y compris dans des contenus multimodaux comme les tableaux [8] et les figures [9], et à construire des benchmarks réalistes pour évaluer ces systèmes dans des conditions proches des usages réels [10].

Méthodes pour le TAL de spécialité

Au-delà de ces deux domaines prioritaires, l'équipe développe des approches plus générales pour le TAL appliqué aux domaines de spécialité. Cela inclut la construction de jeux de données et de benchmarks, l'annotation assistée par modèles de langue, l'usage des LLMs pour l'évaluation, ainsi que l'adaptation ou la conception de modèles capables de mieux prendre en compte les spécificités linguistiques et discursives de documents longs et complexes [3]. Cet axe a vocation à produire des méthodes réutilisables dans d'autres domaines, notamment l'éducation et le droit.

Références

[1] Adrien Bazoge, Pacôme Constant dit Beaufils, Mohammed Hmitouch, Romain Bourcier, Emmanuel Morin, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, and Matilde Karakachoff. Improving social determinants of health documentation in french electronic health records using large language models. *Scientific Reports*, 2025.

[2] Adrien Bazoge, Matthieu Wargny, Pacôme Constant dit Beaufils, Emmanuel Morin, Béatrice Daille, Pierre-Antoine Gourraud, and Samy Hadjadj. Assessing large language models for acute heart failure classification and information extraction from french clinical notes. *Computers in Biology and Medicine*, 195 :110609, 2025.

[3] Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, and Richard Dufour. Is selective masking a key to improving domain adaptation for masked language model ? In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*, ICAIL '25, page 409–413, New York, NY, USA, 2026. Association for Computing Machinery.

[4] Ikram Belmadani, Oumaima El Khettari, Pacôme Constant dit Beaufils, Richard Dufour,

and Benoit Favre. Who judges the judge ? evaluating llm-as-a-judge for french medical open-ended qa. In *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, pages 142–157, 2026.

[5] Ikram Belmadani, Oumaima El Khettari, Benoit Favre, Richard Dufour, et al. Medinjection-fr : Exploring the role of native, synthetic, and translated data in biomedical instruction tuning. In *Language Resources and Evaluation Conference (LREC)*, 2026.

[6] Florian Boudin, Ygor Gallina, and Akiko Aizawa. Keyphrase generation for scientific document retrieval. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1126, Online, July 2020. Association for Computational Linguistics.

[7] Adrien Bougouin, Florian Boudin, and Béatrice Daille. TopicRank : Graph-based topic ranking for keyphrase extraction. In Ruslan Mitkov and Jong C. Park, editors, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

[8] Xanh Ho, Sunisth Kumar, Yun-Ang Wu, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. Table-text alignment : Explaining claim verification against tables in scientific papers. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics : EMNLP 2025*, pages 2509–2517, Suzhou, China, November 2025. Association for Computational Linguistics.

[9] Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. Format matters : The robustness of multimodal llms in reviewing evidence from tables and charts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37) :31014–31022, Mar. 2026.

[10] Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Tian Cheng Xia, Florian Boudin, Andre Greiner-Petter, and Akiko Aizawa. Sciclaimeval : Cross-modal claim verification in scientific papers. In *Proceedings of the Fifteenth biennial Language Resources and Evaluation Conference*, 2026.

[11] Kenichi Iwatsuki, Florian Boudin, and Akiko Aizawa. An evaluation dataset for identifying communicative functions of sentences in English scholarly papers. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France, May 2020. European Language Resources Association.

[12] Leane Jourdan, Nicolas Hernandez, Florian Boudin, and Richard Dufour. Identifying reliable evaluation metrics for scientific text revision. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 6731–6756, Vienna, Austria, July 2025. Association for Computational Linguistics.

[13] Oumaima El Khettari, Virgile Barthet, Guillaume Hocquet, Emmanuel Morin, and Pierre Zweigenbaum. Is clinical text enough ? a multimodal study on mortality prediction in heart failure patients. In *Language Resources and Evaluation Conference (LREC)*, 2026.

[14] Yanis Labrak, Adrien Bazoge, Béatrice Daille, Mickaël Rouvier, and Richard Dufour. How important is tokenization in french medical masked language models ? In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 8223–8234, 2024.

[15] Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. Drbert : A robust pre-trained model in french for biomedical and clinical domains. In *Proceedings*

of the 61st Annual Meeting of the ACL (Volume 1 : Long Papers), pages 16207–16221, 2023.

[16] Yanis Labrak, Adrien Bazoge, Oumaima El Khettari, Mickaël Rouvier, Pacôme Constant Dit Beaufils, Natalia Grabar, Béatrice Daille, Solen Quiniou, Emmanuel Morin, Pierre-Antoine Gourraud, et al. Drbenchmark : A large language understanding evaluation benchmark for french biomedical domain. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 5376–5390, 2024.

[17] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral : A collection of open-source pretrained large language models for medical domains. In *Proceedings of the 62nd Annual Meeting of the ACL (Findings)*, pages 5848–5864, 2024.

[18] Yanis Labrak, Adel Moumen, Richard Dufour, and Mickael Rouvier. Zero-shot end-to-end spoken question answering in medical domain. In *Interspeech*, pages 2020–2024, 2024.

[19] Trung Hieu Ngo, Adrien Bazoge, Solen Quiniou, Pierre-Antoine Gourraud, and Emmanuel Morin. Investigating gender stereotypes in large language models via social determinants of health. In *Findings of the Association for Computational Linguistics (EACL)*, pages 2165–2180, 2026.

AfIA
Association française
pour l'Intelligence Artificielle

Orange Research

Géraldine DAMNATI
geraldine.damnati@orange.com

Yoan CHABOT
yoan.chabot@orange.com

Orange Research
INNOV/DATA&AI

Valentin VIELZEUF
valentin.vielzeuf@orange.com

Elodie GAUTHIER
elodie.gauthier@orange.com

Olivier LE BLOUCH
olivier.leblouch@orange.com

Présentation générale

Les travaux de recherche menés à Orange autour du traitement automatique des langues (TAL) et des connaissances sont structurés en différents projets dont les thématiques sont décrites dans la section suivante. Nous menons des travaux de recherche amont et des travaux guidés par les besoins applicatifs du Groupe Orange, en maintenant un équilibre entre nos liens avec les entités opérationnelles et la recherche ouverte et collaborative. Nous collaborons avec plusieurs partenaires académiques au niveau national et international et le co-encadrement des thèses en cours est assuré par le LIS (Aix Marseille Université), le LIUM (Le Mans Université), le LORIA (Université de Lorraine) et EURECOM.

Langage

Une partie significative de nos travaux actuels est consacrée à l'adaptation des modèles de langage à un domaine, avec un travail sur l'adaptation au domaine technique des Telecoms [7] et la participation à un challenge organisé par l'AMIAD sur l'adaptation au domaine de la défense [40]. Ces travaux trouvent actuellement un cadre plus large à travers le projet européen LLMs4EU où Orange coordonne les travaux liés au domaine applicatif des Telecoms et s'articulent autour des besoins opérationnels et les cas d'usage du Groupe Orange. L'adaptation des modèles de langues à des langues peu dotées est également un sujet d'étude, avec plus particulièrement l'adaptation à des langues africaines parlées dans les pays où Orange est opérateur.

Dans le but d'approfondir notre compréhension des mécanismes sous-jacents aux modèles de langage de type Transformer, des travaux plus amont sont menés d'un côté sur les notions de connaissances [3, 2] et leur évolution temporelle [35] et d'autre part sur la notion de tâches avec des méthodes statistiques cherchant à caractériser des relations d'inclusion [20] et des méthodes de combinaison de tâches [43].

Nous menons également des travaux de recherche dans le domaine élargi de l'IA agentique. Un premier axe de recherche est dédié aux aspects dialogiques des agents conversationnels dans un contexte d'interactions orientées tâche, avec des travaux en lien avec les méthodes agentiques [19], la production de ressources [9], l'évaluation et la simulation dans le contexte des agents conversationnels [25] et la notion d'*entrainment* [34]. Une thèse est également en cours sur l'évaluation des agents conversationnels [18] et la notion de grounding . D'autres travaux s'intéressent à la personnalisation et aux comportements sociaux [44]. Enfin, une thèse a démarré fin 2025 sur les modèles de raisonnement dans le contexte du *Deep Research*.

L'évaluation occupe également une place centrale dans nos travaux de recherche, avec par exemple des travaux collaboratifs sur l'évaluation des modèles génératifs guidée par un contrôle sémantique [4], et la participation au projet ANR-AAPG2025 PRCI franco-canadien REGARD (*Ro-*

bust Evaluation of Generative AI in Real-World Environments).

La multimodalité constitue un axe de recherche en développement, avec par exemple la participation au projet MINERAL, visant à étudier les représentations multimodales des conversations ; ainsi que nos travaux sur l'interprétation des modèles génératifs de vision et de langage [10].

Enfin, la linguistique computationnelle demeure un axe de recherche structurant et nous poursuivons nos recherches en analyse syntaxique et sémantique avec un accent plus particulier ces derniers temps sur l'analyse AMR (Abstract Meaning Representation) [28, 27, 21], ainsi qu'une participation au challenge dédié à l'analyse UMR [26]. Une thèse a commencé début 2026 sur l'analyse AMR appliquée aux conversations.

Connaissances

Les graphes de connaissances constituent un outil central pour structurer et exploiter les données hétérogènes générées dans le secteur des télécommunications. Dans ce cadre, Orange mène depuis plusieurs années des travaux sur la construction, l'évolution et la maintenance de graphes de connaissances d'entreprise, en s'appuyant sur des sources variées.

Parmi les différents types de données pouvant être utilisés pour construire un graphe de connaissances, les documents textuels représentent une source d'information importante [39, 41] mais nos équipes s'intéressent également à l'extraction d'informations à partir de données tabulaires [36, 29] et de figures [8]. Nous étudions également des méthodes traitant conjointement plusieurs modalités comme dans le projet ANR ECLADATTA considérant les tables et le texte.

Un enjeu central de nos travaux concerne la qualité des graphes de connaissances, en particulier face à l'hétérogénéité des sources. Nous menons des recherches sur l'évaluation de la qualité et les méthodes de réconciliation permettant de gérer les contradictions, les différences terminologiques et les variations de spécificité entre sources. Ces travaux comprennent un état de l'art sur la gestion de l'incertitude dans les graphes de connaissances [31] et l'ensemble d'outils et de jeux de données trustfuse pour réaliser des opérations de fusion de connaissances [33, 32].

Nos travaux d'ingénierie des connaissances s'étendent également à la modélisation de verticales d'intérêt pour le Groupe Orange. Nous mentionnerons à titre d'exemple l'ontologie open-source NORIA-O [42], un modèle des infrastructures réseaux et de leurs activités à grande échelle.

Parole

La parole constitue un axe important de nos travaux en TAL, à l'interface entre reconnaissance automatique, apprentissage de représentations et modélisation sémantique. Nos recherches portent à la fois sur la performance des systèmes, leur coût de calcul et leur capacité à intégrer des informations de plus haut niveau.

Un premier volet, historiquement central pour Orange, concerne la reconnaissance automatique de la parole, avec plus récemment des travaux sur le streaming. Dans ce cadre, nous étudions l'impact des choix d'architecture et d'entraînement sur les performances en temps réel, en particulier sous fortes contraintes de latence [14, 15]. Ces travaux s'inscrivent dans une réflexion plus générale sur l'adaptation des modèles de reconnaissance aux usages contraints.

Un deuxième volet porte sur les approches auto-supervisées pour la parole et la réduction de complexité. Nous cherchons à entraîner ce type de modèles avec un budget de calcul maîtrisé, tout en préservant un niveau de qualité compatible avec les usages visés [38, 37]. Cet axe répond à un enjeu important de sobriété computationnelle dans le développement des modèles de parole.

Enfin, nous menons des travaux à l'interface entre parole et sémantique, portant notamment sur le spoken dialogue state tracking avec notre solution olisia[30], sur la structuration sémantique de la parole dialogique [16], ainsi que sur les méthodes d'annotation adaptées à ces données [17]. Nous nous intéressons également aux approches récentes de type SpeechLLM [24] et aux sentence-level embeddings, en particulier pour analyser leurs propriétés et leur apport dans des tâches de compréhension de la parole [1].

Langues africaines

Un projet est dédié au Traitement Automatique de la Parole pour les langues peu dotées, et en particulier les langues africaines. Les travaux récents ont porté sur la constitution de ressources linguistiques [22] et sur le développement de modèles auto-supervisés [12, 11] pour la réalisation de tâches avales comme la reconnaissance automatique de la parole ou l'identification des langues parlées, et ce pour diverses langues africaines.

En 2025, une extension des modèles de représentation de la parole de type HuBERT, spécifiquement pré-entraînés sur des enregistrements vocaux issus de divers pays du continent africain, a été réalisée vers des architectures large (300 millions de paramètres) et extra-large (900 millions de paramètres) [13]. Les modèles sont disponibles [9] et les résultats ont été présentés lors d'une journée d'étude AFIA-ATALA sur les technologies linguistiques pour les langues peu dotées.

Nous contribuons également à plusieurs projets de recherche partenariale, notamment par l'accompagnement scientifique du projet SunuChat porté par l'École Polytechnique de Thiès (Sénégal) pour un assistant conversationnel écrit et vocal, en wolof et en français, sachant répondre à des questions relatives aux thématiques médicales et agricoles). Nos travaux ont été mis en avant lors de la tenue de la Conférence mondiale sur le développement des pays à faibles et moyens revenus [10], organisée par le Global Development Network.

Une thèse consacrée à la compréhension automatique de la parole pour les langues parlées en Afrique subsaharienne, a démarré fin 2025.

Voix

En complément des travaux sur la parole, nous étudions également les attributs de la voix dans le but de développer des systèmes de génération et de conversion de voix contrôlables. Nous cherchons à démêler [45], extraire et contrôler des attributs vocaux liés notamment au timbre, à l'accent et à la paralinguistique. Ces travaux sont associés au projet ANR EVA coordonné par Orange.

Un premier axe concerne le développement de modèles permettant de démêler les caractéristiques vocales en deux composantes : le timbre (qualité intrinsèque liée aux caractéristiques biologiques) et le non-timbre (prosodie, rythme, accent, émotion) [23]. Ces modèles ont été rendus publics sur Hugging Face et sont utilisés dans des applications de synthèse vocale contrôlée.

Un second volet de recherche, cadré par le Voice Privacy Challenge, porte sur l'anonymisation de la voix et en particulier sur l'impact des attributs non-timbraux (prosodie, rythme, accent) dans la protection de l'identité vocale. Il a été montré que ces attributs peuvent permettre d'identifier un locuteur même après anonymisation [5]. Des méthodes d'identification de l'accent ont également été développées [6] dans le but d'analyser le taux de préservation de l'accent après anonymisation.

Enfin, nous démarrons des travaux sur la détection de deepfakes audio via une thèse de façon à capitaliser sur notre expertise en matière d'IA générative et d'IA de confiance.

Membres Permanents

- Pierre ADAM (Data Scientist)
- Elys ALLESIARDO (Data Scientist)
- Munshi ASADULLAH (Chercheur)
- Quentin BRABANT (Data scientist)
- Antoine CAUBRIÈRE (Chercheur)
- Yoan CHABOT (Chercheur)
- Géraldine DAMNATI (Chercheuse)
- Elodie GAUTHIER (Chercheuse)
- Nicolas GENGEMBRE (Chercheur)
- Johannes HEINECKE (Chercheur)
- Léane JOURDAN (Chercheuse)
- Charbel KINDJI (Data Scientist)
- Olivier LE BLOUCH (Chercheur)
- Gwénolé LECORVÉ (Chercheur, HDR)
- Emmanuel MORY (Responsable)
- Ghislain PUTOIS (Data Scientist)
- Lina ROJAS-BARAHONA (Chercheuse, HDR)
- Constance SCHERRER (Data Scientist)
- Anastasia SHIMORINA (Data Scientist)
- Lionel TAILHARDAT (Chercheur)
- Morgan VEYRET (Ingénieur logiciel)
- Valentin VIELZEUF (Chercheur)

9. https://huggingface.co/collections/Orange/african-speech-foundation-models
10. https://www.gdn.int/global-development-conference-2025

- Benjamin WHITE (Data Scientist)
- Olivier ZHANG (Data Scientist)

Références

[1] Elys Allesiardo, Antoine Caubrière, and Valentin Vielzeuf. Forewarned is forearmed : when non-sequential embedding turns into an anomaly detector. In *LREC*, 2026.

[2] Hichem Ammar Khodja, Abderrahmane Ait gueni ssaid, Frederic Bechet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. Factual Knowledge Assessment of Language Models Using Distractors. In *COLING*, 2025.

[3] Hichem Ammar Khodja, Frédéric Béchet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. Wikifactdiff : A Large, Realistic, and Temporally Adaptable Dataset for Atomic Factual Knowledge Update in Causal Language Models. In *LREC-COLING*, 2024.

[4] Elie Antoine, Frederic Bechet, Géraldine Damnati, and Philippe Langlais. A linguistically-motivated evaluation methodology for unraveling model's abilities in reading comprehension tasks. In *EMNLP*, 2024.

[5] Rayane Bakari, Olivier Le Blouch, Nicholas Evans, Nicolas Gengembre, Michele Panariello, and Massimiliano Todisco. The influence of non-timbral cues in voice anonymisation and evaluation. In *5th Symposium on Security and Privacy in Speech Communication*, 2025.

[6] Rayane Bakari, Olivier Le Blouch, Nicolas Gengembre, Nicholas Evans, and Michele Panariello. Identity leakage through accent cues in voice anonymisation. *arXiv preprint arXiv :2603.26843*, 2026.

[7] Camille Barboule, Viet-Phi Huynh, Adrien Bufort, Yoan Chabot, Géraldine Damnati, and Gwénolé Lecorvé. Telcolm : collecting data, adapting, and benchmarking language models for the telecommunication domain. *arXiv preprint arXiv :2412.15891*, 2024.

[8] Camille Barboule, Benjamin Piwowarski, and Yoan Chabot. Survey on question answering over visually rich documents : Methods, challenges, and trends. *arXiv preprint arXiv :2501.02235*, 2025.

[9] Quentin Brabant, Rojas-Barahona Lina, Gwénolé Lecorvé, and Claire Gardent. KGConv, a conversational corpus grounded in wikidata. In *LREC-COLING*, 2024.

[10] Michele Cafagna, Lina M. Rojas-Barahona, Kees van Deemter, and Albert Gatt. Interpreting vision and language generative models with semantic visual priors. *Frontiers in Artificial Intelligence*, Volume 6, 2023.

[11] Antoine Caubrière and Elodie Gauthier. Africa-centric self-supervised pretraining for multilingual speech representation in a sub-saharan context. In *5th Workshop on African Natural Language Processing*, 2024.

[12] Antoine Caubrière and Elodie Gauthier. Représentation de la parole multilingue par apprentissage auto-supervisé dans un contexte subsaharien. In *JEP-TALN*, 2024.

[13] Antoine Caubrière and Elodie Gauthier. Scaling hubert for african languages : From base to large and xl. *arXiv preprint arXiv :2511.23370*, 2025.

[14] Youness Dkhissi, Valentin Vielzeuf, Elys Allesiardo, and Anthony Larcher. Do we really need self-attention for streaming automatic speech recognition ? In *ICASSP*, 2026.

[15] Youness Dkhissi, Valentin Vielzeuf, Elys Allesiardo, and Anthony Larcher. Sens-asr : Semantic embedding injection in neural-transducer for streaming automatic speech recognition. In *LREC*, 2026.

[16] Lucas Druart. *Vers une Compréhension Contextuelle et Structurée de la Parole Dialogique Orientée Tâche*. PhD thesis, Université d'Avignon, 2024.

[17] Lucas Druart, Valentin Vielzeuf, and Yannick Estève. Investigating low-cost llm annotation for spoken dialogue understanding datasets. In *TSD*, 2024.

[18] Michelle Elizabeth, Alicja Kasicka, Natalia Krawczyk, Magalie Ochs, Gwénolé Lecorvé, Justyna Gromada, and Lina M Rojas-Barahona. Neural models and language model

prompting for the multidimensional evaluation of open-ended conversations. In *12th DST Challenge*, 2025.

[19] Michelle Elizabeth, Morgan Veyret, Miguel Couceiro, Ondrej Dusek, and Lina M Rojas-Barahona. Exploring ReAct Prompting for Task-Oriented Dialogue : Insights and Shortcomings. In *IWSDS*, 2025.

[20] Loïc Fosse, Frederic Bechet, Benoit Favre, Géraldine Damnati, Gwénolé Lecorvé, Maxime Darrin, Philippe Formont, and Pablo Piantanida. Statistical Deficiency for Task Inclusion Estimation. In *ACL*, 2025.

[21] Venera Gareeva and Johannes Heinecke. Modelling idiomatic expressions in abstract meaning representation. In *Workshop on Designing Meaning Representations @ LREC 2026*, 2026.

[22] Elodie Gauthier, Aminata Ndiaye, and Abdoulaye Guissé. Kallaama : A transcribed speech dataset about agriculture in the three most widely spoken languages in Senegal. In Rooweither Mabuya, Muzi Matfunjwa, Mmasibidi Setaka, and Menno van Zaanen, editors, *Workshop on Resources for African Indigenous Languages @ LREC-COLING*, May 2024.

[23] Nicolas Gengembre, Olivier Le Blouch, and Cédric Gendrot. Disentangling prosody and timbre embeddings via voice conversion. In *INTERSPEECH*, 2024.

[24] Nizar El Ghazal, Antoine Caubrière, and Valentin Vielzeuf. The speech-llm takes it all : A truly fully end-to-end spoken dialogue state tracking approach. In *LREC*, 2026.

[25] Justyna Gromada, Alicja Kasicka, Ewa Komkowska, Lukasz Krajewski, Natalia Krawczyk, Morgan Veyret, Bartosz Przybył, Lina M Rojas-Barahona, and Michał K Szczerbak. Evaluating conversational agents with persona-driven user simulations based on large language models : A sales bot case study. In *EMNLP : Industry Track*, 2025.

[26] Johannes Heinecke and Munshi Asadullah. Orange @ UMR Parsing Shared Task. In *Workshop on Designing Meaning Representations @ LREC 2026*, 2026.

[27] Johannes Heinecke, Munshi Asadullah, Frédéric Herledan, and Geraldine Damnati. Creating a high quality abstract meaning representation dataset automatically. In *LREC*, 2026.

[28] Johannes Heinecke, Maria Boritchev, and Frédéric Herledan. Question-parsing with abstract meaning representation enhanced by adding small datasets. In *NoDaLiDa/Baltic-HLT*, 2025.

[29] Viet-Phi Huynh, Yoan Chabot, and Raphaël Troncy. Towards generative semantic table interpretation. In *VLDB Workshops*, 2023.

[30] Léo Jacqmin, Lucas Druart, Yannick Estève, Benoît Favre, Lina M Rojas-Barahona, and Valentin Vielzeuf. Olisia : a cascade system for spoken dialogue state tracking. In *11th DST Challenge*, 2023.

[31] Lucas Jarnac, Yoan Chabot, and Miguel Couceiro. Uncertainty management in the construction of knowledge graphs : a survey. *arXiv preprint arXiv :2405.16929*, 2024.

[32] Lucas Jarnac, Yoan Chabot, and Miguel Couceiro. Trustfuse : A fusion testbed for uncertain knowledge reconciliation. In *KCAP*, 2025.

[33] Lucas Jarnac, Yoan Chabot, and Miguel Couceiro. Wikiconflict : A new dataset for conflicting data reconciliation in knowledge graph construction. In *KCAP*, 2025.

[34] Jay Kejriwal, Štefan Beňuš, and Lina M. Rojas-Barahona. Entrainment detection using DNN. *Computer Speech & Language*, 99 :101930, 2026.

[35] Hichem Ammar Khodja, Frederic Bechet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. Factual Knowledge in Language Models : Robustness and Anomalies under Simple Temporal Context Variations. In *ACL-Workshop on Large Language Model Memorization (L2M2)*, 2025.

[36] Jixiong Liu, Yoan Chabot, Raphaël Troncy, Viet-Phi Huynh, Thomas Labbé, and Pierre Monnin. From tabular data to knowledge graphs : A survey of semantic table interpretation tasks and methods. *Journal of Web Semantics*, 76 :100761, 2023.

[37] Luis Lugo and Valentin Vielzeuf. Efficiency-oriented approaches for self-supervised speech representation learning. *International Journal of Speech Technology*, 27(3) :765–779, 2024.

[38] Luis Lugo and Valentin Vielzeuf. Towards efficient self-supervised representation learning in speech processing. In *Findings EACL*, 2024.

[39] Carmelle Meli Songuon, Yoan Chabot, Lucas Jarnac, and Viet-Phi Huynh. A guided and flexible llm-based approach for knowledge extraction from text. In *KCAP*, 2025.

[40] Ismaël Rousseau, Claire Perroux, Pierre Adam, Thomas Girault, Lionel Delphin-Poulat, Morgan Veyret, Gwénolé Lecorvé, and Géraldine Damnati. O_FT@ EvalLLM2025 : étude comparative de choix de données et de stratégies d'apprentissage pour l'adaptation de modèles de langue à un domaine. In *challenge EvalLLM@CORIA-TALN*, 2025.

[41] Anastasia Shimorina, Johannes Heinecke, and Frédéric Herledan. Knowledge extraction from texts based on wikidata. In *NAACL-HLT : Industry Track*, 2022.

[42] Lionel Tailhardat, Yoan Chabot, and Raphaël Troncy. Noria-o : an ontology for anomaly detection and incident management in ict systems. In *ESWC*, 2024.

[43] Brahim Touayouch, Loïc Fosse, Géraldine Damnati, and Gwénolé Lecorvé. DivMerge : A divergence-based model merging method for multi-tasking. In *EACL*, 2026.

[44] Benjamin White and Anastasia Shimorina. Predicting Social Media User Actions : A Hybrid Approach for Common and Rare Behavior Prediction on Bluesky. In *Workshop on Social Context (SoCon) and Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ LREC 2026*, 2026.

[45] Olivier Zhang, Olivier Le Blouch, Nicolas Gengembre, and Damien Lolive. An extension of disentanglement metrics and its application to voice. In *Interspeech*, 2023.

AfIA
Association française
pour l'Intelligence Artificielle

TETIS/GeoTexIA : GÉOinformations, données TEXtuelles et Intelligence Artificielle

UMR TETIS
CIRAD, INRAE, AgroParisTech, CNRS
https://umr-tetis.cirad.fr/organisation/groupes-de-recherche/gr-geotexia

Rémy DECOUPES
remy.decoupes@inrae.fr

Mathieu ROCHE
mathieu.roche@cirad.fr

Présentation générale

GeoTexIA (GÉOinformations, données TEXtuelles et Intelligence Artificielle) de l'UMR TETIS est un groupe de recherche consacré à l'extraction et à la mise en relation d'informations géographiques et de connaissances issues de textes grâce à des méthodes d'Intelligence Artificielle (IA). Le groupe conçoit et développe des approches permettant d'extraire, de structurer et de croiser des connaissances provenant de sources hétérogènes, souvent non structurées, telles que les textes, les images satellitaires, les capteurs in situ ou les données issues de divers modèles.

Les disciplines mobilisées par l'équipe relèvent du traitement automatique de la langue, de la fouille de textes et, plus largement, de données ainsi que du web sémantique. Les cas d'étude sont l'épidémiologie, la sécurité alimentaire, le suivi de crises, mais aussi les transitions agroécologiques ou le suivi des territoires. Le groupe développe notamment la plateforme PADI-web [11], dédiée à la détection précoce de l'émergence et de la propagation d'épidémies dans un contexte *One Health*.

Les sources de données textuelles sont issues d'articles de presse, de publications de réseaux sociaux, d'articles scientifiques, de documents juridiques ainsi que des données audio provenant notamment de radios locales et de contenus diffusés sur YouTube.

Enfin, notre approche est structurée autour de l'information géographique, utilisée à la fois comme cadre d'intégration des différentes sources de données et comme objet d'étude, notamment à travers l'analyse des biais géographiques présents dans les modèles de langage.

11. https ://www.padi-web-one-health.org

Principaux thèmes de recherche

Les activités de GeoTexIA en traitement automatique des langues s'organisent autour de quatre axes principaux :

Veille événementielle et recherche d'information

Le groupe développe des méthodes d'extraction d'évènements et de recherche d'information dans des flux multilingues. Ces travaux concernent notamment l'identification précoce d'évènements épidémiologiques [12, 13, 8], ou encore en lien avec la sécurité alimentaire [9] à partir de données issues de la presse. D'autres activités en cours s'intéressent à la détection de signaux faibles qui pourraient présager l'apparition d'un évènement ou bien encore à des méthodologies visant à supprimer des contenus de désinformation [5]. Les approches développées utilisent, adaptent, combinent et étendent des modèles de langues pour annoter et augmenter les données de domaines spécialisés.

Information géographique et fouille de texte

Cet axe vise l'identification, la désambiguïsation [4] et la représentation des entités géographiques et des relations spatiales présentes dans les textes. Les recherches portent sur l'extraction d'informations géographiques, leur intégration dans des graphes de connaissances [9] ainsi que leur exploitation pour l'analyse spatiale et la cartographie d'informations textuelles. L'équipe propose également des méthodologies sur le géocoding, tâche qui consiste à attribuer des coordonnées géographiques à des mentions de lieux, d'entités spatiales relatives [11] comme *dans le nord de Paris*.

Biais géographique et explicabilité des modèles de langues

Les travaux récents du groupe étudient les représentations géographiques apprises par les modèles de langage. Ils visent à caractériser les biais spatiaux présents dans les données d'entraînement et dans les prédictions des modèles [2], ainsi qu'à développer des méthodes d'évaluation et d'explicabilité.

Intégration de données hétérogènes

Le groupe développe des approches d'intégration combinant des données textuelles, géographiques, tabulaires, des images et des séries temporelles. Ces travaux mobilisent les graphes de connaissances, les technologies du web sémantique et les approches multimodales afin de produire des représentations unifiées de phénomènes complexes observés à différentes échelles spatiales et temporelles. Les méthodes développées sont également évaluées sur des domaines appliqués comme les facteurs favorisant la propagation d'une maladie [1] ou bien encore la prédiction du niveau de sécurité alimentaire à différentes échelles [3]. Enfin, l'équipe mène des travaux qui visent à mettre en lien différents champs de bases de données [6].

Plateformes

PADI-web pour la santé animale et santé végétale

PADI-web est une plateforme de veille épidémiologique exploitant des techniques de fouille de textes et d'extraction d'information pour détecter précocement des évènements sanitaires à partir de sources ouvertes. Initialement développée pour la santé animale [12, 13], la plateforme est aujourd'hui adaptée à la santé végétale [10] et publique.

Production de données textuelles

Le groupe produit de nombreux corpus originaux en anglais, français et en langue arabe qui sont publiés dans des plateformes institutionnelles :

- https://github.com/mehtab-alam/GeospaCy
- https://dataverse.cirad.fr/dataverse/tetis

Production logicielle et science ouverte

Le groupe développe également plusieurs ressources logicielles dédiées à l'extraction d'entités géographiques et à l'analyse spatiale de corpus textuels notamment GeoSpaCy [7] et SNEToolkit [4]. La liste de nos productions logicielles est diponible en ligne : https://github.com/tetis-nlp.

Membres Impliqués

- Rémy DECOUPES (IR, Inrae)
- Mathieu ROCHE (DR, CIRAD)
- Landy RAJAONARIVO (CR, Inrae)
- Maguelonne TEISSEIRE (DR, Inrae)
- Annelise TRAN (DR, CIRAD)
- Sarah VALENTIN (CR, CIRAD)

Références

[1] Bahdja Boudoua, Mathieu Roche, Maguelonne Teisseire, and Annelise Tran. EpiDCA : Adaptation and implementation of a danger theory algorithm for event-based epidemiological surveillance. *Computers and Electronics in Agriculture*, 229 :109693, February 2025.

[2] Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. Evaluation of geographical distortions in language models. *Machine Learning*, 114(12) :263, December 2025.

[3] Hugo Deléglise, Roberto Interdonato, Agnès Bégué, Elodie Maître d'Hôtel, Maguelonne Teisseire, and Mathieu Roche. Food security prediction from heterogeneous data combining machine and deep learning methods. *Expert Systems with Applications*, 190 :116189, March 2022.

[4] Rodrique Kafando, Rémy Decoupes, Mathieu Roche, and Maguelonne Teisseire. SNEToolkit : Spatial named entities disambiguation toolkit. *SoftwareX*, 23 :101480, July 2023.

[5] Youssef Mahdoubi, Najlae Idrissi, Mathieu Roche, and Sarah Valentin. Enhancing language models with selective masking for thematic and misinformation classification in a one health context. *Expert Systems with Applications*, 322 :132345, 2026.

[6] Oussama Mechhour, Sandrine Auzoux, Clément Jonquet, and Mathieu Roche. Maeva : A hybrid approach for matching agroecological experiment variables. *Natural Language Processing Journal*, 13 :100180, 2025.

[7] Syed Mehtab Alam, Elena Arsevska, Mathieu Roche, and Maguelonne Teisseire. GeospaCy : A tool for extraction and geographical referencing of spatial expressions in textual data. In Nikolaos Aletras and Orphee De Clercq, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, pages 115–126, St. Julians, Malta, March 2024. Association for Computational Linguistics.

[8] Edmond Menya, Roberto Interdonato, Dickson Owuor, and Mathieu Roche. Novelty Detection in Event Surveillance Documents. *IEEE Access*, 14 :29566–29589, 2026.

[9] C. A. Ngom, L. Rajaonarivo, M. Teisseire, and S. Valentin. Spatio-temporal knowledge graph from unstructured texts : A multi-scale approach for food security monitoring. *AGILE : GIScience Series*, 7 :16, 2026.

[10] Mathieu Roche, Julien Rabatel, Carlène Trevennec, and Isabelle Pieretti. Padi-web for plant health surveillance. In Shareeful Islam and Arnon Sturm, editors, *Intelligent Information Systems*, pages 148–156, Cham, 2024. Springer Nature Switzerland.

[11] Mehtab Alam Syed, Elena Arsevska, Mathieu Roche, and Maguelonne Teisseire. GeospatRE : extraction and geocoding of spatial relation entities in textual documents. *Cartography and Geographic Information Science*, 52(3) :221–236, May 2025.

[12] Sarah Valentin, Elena Arsevska, Sylvain Falala, Jocelyn de Goër, Renaud Lancelot, Alizé Mercier, Julien Rabatel, and Mathieu Roche. PADI-web : A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169 :105163, February 2020.

[13] Sarah Valentin, Elena Arsevska, Julien Rabatel, Sylvain Falala, Alizé Mercier, Renaud Lancelot, and Mathieu Roche. PADI-web 3.0 : A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health*, page 100357, December 2021.