



CORIA-TALN 2025

*20e Conférence en Recherche d'Information et Applications (CORIA)
32ème Conférence sur le Traitement Automatique des Langues
Naturelles (TALN)
27ème Rencontre des Étudiants Chercheurs en Informatique pour le
Traitement Automatique des Langues (RECITAL)
Les 18e Rencontres Jeunes Chercheurs en RI (RJCRI)
(CORIA-TALN)¹*

Actes de CORIA-TALN-RJCRI-RECITAL 2025.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles
(TALN),
volume 1 : articles scientifiques originaux

Frédéric BECHET, Adrian-Gabriel CHIFU, Karen PINEL-SAUVAGNAT, Benoit FAVRE, Eliot MAES,
Diana NURBAKOVA (Éds.)

Marseille, France, 30 juin au 4 juillet 2025

1. <https://coria-taln-2025.lis-lab.fr>

Avec le soutien de

Organisateurs



Soutiens académiques



Sponsors privés



Préface

Organisée par l’Université d’Aix-Marseille et les UMR CNRS LIS et LPL, sous l’égide de l’Association francophone de Recherche d’Information et Applications (ARIA) et l’Association pour le Traitement Automatique des Langues (ATALA), l’édition 2025 de CORIA-TALN regroupe :

- la 20e Conférence en Recherche d’Information et Applications (CORIA) ;
- la 32e Conférence sur le Traitement Automatique des Langues Naturelles (TALN) ;
- les 27e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) ;
- les 18e Rencontres Jeunes Chercheurs en RI (RJCRI).

Les conférences CORIA et TALN offrent le plus important forum d’échange francophone aux acteurs universitaires et industriels des technologies de la langue et la recherche d’information. Pour cette édition, nous avons plus de 330 inscrits dont une grande partie des étudiants qui construisent le futur de la recherche francophone et assurent le relais de son développement. La conférence principale compte plus de 260 inscrits et les ateliers plus de 70 inscrits.

Les conférencières et conférenciers invités de la conférence sont Marine CARPUAT, de l’Université du Maryland (USA), Mohamed CHETOUANI de Sorbonne Université, et Owen RAMBOW de Stony Brook University (USA). Ces invités représentent un large spectre de thématiques dans le domaine de la recherche d’information et le traitement automatique des langues, et partageront les dernières avancées dans leur domaine d’expertise.

En termes de soumissions, 101 articles ont été soumis à TALN, 76 acceptés dont 51 contributions scientifiques originales et 25 traductions d’articles déjà publiés dans des conférences internationales majeures du domaine. Il n’y a pas eu de différenciation dans le processus de sélection entre oraux et posters. Un total de 29 articles a été soumis à CORIA dont 25 ont été retenus pour présentation (8 courts, 13 longs et 4 résumés). RECITAL-RJCRI a accepté 17 articles sur 19 soumissions. En complément de la conférence principale, huit ateliers sont présentés : Avancement de l’AMR et de l’Analyse Sémantique (4AS), Accès à l’information basé sur le dialogue et grands modèles de langage (DIAG-LLM), Traitement de données langagières dynamiques par les outils et méthodes du TAL (DYN-TAL), Ethic and Alignment of (Large) Language Models (EALM), Évaluation des modèles génératifs (LLM) et challenge (EvalLLM), Intelligence Artificielle générative et ÉDUcation : Enjeux, Défis et Perspectives de Recherche (IA-ÉDU), Traitement du langage médical à l’époque des LLMs (MLP-LLM) et Science Participative pour les Données et Corpus Linguistiques (ParCoL). Ce programme est complété par une session industrielle ciblant explicitement des contributions qui montrent le développement des technologies associées au langage dans l’industrie. Ces événements illustrent à la fois des tendances nouvelles présentes dans la communauté et des activités récurrentes.

Il convient d’exprimer une profonde reconnaissance envers toutes les personnes qui ont participé à faire vivre la conférence, d’un côté les auteurs de toutes les soumissions et de l’autre les membres de différents comités scientifiques de la conférence. Un remerciement très chaleureux aux relecteurs qui ont accepté une charge importante et qui ont fait des relectures d’urgence afin de faciliter le bon déroulement de la conférence. La bienveillance et l’expertise des comités de programme ont permis la constitution d’un programme riche en thématiques et d’un niveau scientifique correspondant aux attentes de la communauté. Il est également essentiel d’exprimer notre gratitude envers les sponsors et les organisations qui ont subventionné la conférence. Leur soutien financier a permis à cet événement scientifique de se réaliser dans les meilleures conditions, rappelant l’importance des aspects financiers dans la réussite de telles initiatives. Finalement, un grand merci aux différentes équipes présentes pour le bon fonctionnement, notamment des équipes de l’ATALA et de l’ARIA qui nous ont accompagnés dans les différentes étapes

de l'organisation.

- Frédéric BÉCHET et Adrian-Gabriel CHIFU, présidents du comité d'organisation de CORIA-TALN
- Karen PINEL-SAUVAGNAT, présidente du comité de programme de CORIA
- Benoit FAVRE, président du comité de programme de TALN
- Eliot MAES et Diana NURBAKOVA, présidents du comité de programme de RECITAL-RJCRI

Comités

Comité de Programme

- Benoit FAVRE (Aix Marseille Université – LIS)
- Frederic BECHET (Aix Marseille Université – LIS)
- Marie CANDITO (Université Paris 7 / INRIA)
- Vincent CLAVEAU (Irisa)
- Caio CORRO (Université Paris-Saclay)
- Benoît CRABBÉ (Paris 7 et INRIA)
- Iris ESHKOL-TARAVELLA (University of Orléans)
- Karèn FORT (Sorbonne Univers)
- Natalia GRABAR (STL CNRS Université Lille 3)
- Thierry HAMON (Université Paris-Saclay, CNRS, LIMSI, Université Sorbonne Paris Nord)
- Lydia-Mai HO-DAC (CLLE)
- Philippe LANGLAIS (University of Montreal)
- Luce LEFEUVRE (DTIPG, SNCF)
- José MORENO (IRIT/UPS)
- Aurélie NÉVÉOL (CNRS)
- François PORTET (Laboratoire d’Informatique de Grenoble)
- Solen QUINIOU (LS2N – Nantes Université)

Comité de Relecture

- Eunice AKANI (Aix-Marseille Université, Programme CEDRE – IDEAL France 2030)
- Alexandre ALLAUZEN (Paris-Dauphine Université, ESPCI, PSL)
- Maxime AMBLARD (Université de Lorraine)
- Hichem AMMAR KHODJA (Orange)
- Elie ANTOINE (Aix-Marseille Université)
- Delphine BATTISTELLI (MoDyCo- CNRS Université Paris Nanterre)
- Patrice BELLOT (Aix-Marseille Université – CNRS (LIS))
- Asma BEN ABACHA (Microsoft)
- Farah BENAMARA (Univ. Paul Sabatier, Toulouse and IPAL, Singapore)
- Timothée BERNARD (Université Paris Cité)
- Sam BIGEARD (INRIA Nancy, France)
- Florian BOUDIN (Université de Nantes)
- Maxime BOUTHORS (Sorbonne Université, CNRS, ISIR)
- Quentin BRABANT (Orange Labs, Lannion, France)
- Chloé BRAUD (IRIT – CNRS)
- François BUET (Université de Lorraine, LORIA)
- Nicolas BÉCHET (IRISA)
- Remi CARDON (CENTAL, IL&C, Université Catholique de Louvain)
- Christophe CERISARA (LORIA)
- Adrian-Gabriel CHIFU (Aix Marseille Univ, Université de Toulon, CNRS, LIS)
- Thibault CLÉRICE (ALMANaCH, Inria)
- Maximin COAVOUX (CNRS, Université Grenoble Alpes)
- Josep CREGO (ChapsVision)
- Béatrice DAILLE (Laboratoire d’Informatique Nantes Atlantique (LINA))
- Rodolfo DELMONTE (Universita’ Ca’ Foscari)
- Gaël DIAS (Normandie University)

- Marco DINARELLI (LIG)
- Fanny DUCCEL (Université Paris-Saclay, LISN)
- Richard DUFOUR (LS2N – Nantes University)
- Yoann DUPONT (Lattice, Sorbonne Nouvelle)
- Olivier FERRET (CEA List)
- Thierry FONTENELLE (European Investment Bank)
- Amel FRAISSE (Université de Lille)
- Thomas FRANCOIS (Université catholique de Louvain)
- Lingyun GAO (Université Catholique de Louvain)
- Barbara GENDRON (LORIA – University of Lorraine)
- Thomas GERALD (Université Paris Saclay, CNRS, SATT, LISN)
- Sahar GHANNAY (LISN lab)
- Carlos-Emiliano GONZÁLEZ-GALLARDO (University of Tours)
- Loïc GROBOL (MoDyCo)
- Cyril GROUIN (Université Paris-Saclay, CNRS, LISN)
- Gaël GUIBON (Université de Lorraine – LORIA)
- Vincent GUIGUE (Agroparistech)
- Camille GUINAUDEAU (Université Paris-Saclay, CNRS, JFLI, 101-0003 Tokyo, Japon)
- Nabil HATHOUT (CNRS)
- Nicolas HERNANDEZ (Nantes Université – LS2N CNRS UMR 6004)
- Nicolas HERVÉ (INA)
- Stéphane HUET (LIA – Université d’Avignon)
- Mélanie JOUITTEAU (UMR 5478, CNRS)
- Léane JOURDAN (Nantes University)
- Cyril LABBÉ (Univ. Grenoble Alpes)
- Mathieu LAFOURCADE (LIRMM)
- Guy LAPALME (University of Montreal)
- Thomas LAVERGNE (LIMSI, CNRS, Univ. Paris Sud, Université Paris Saclay)
- Gwénolé LECORVÉ (Orange)
- Benjamin LECOUTEUX (Laboratoire Informatique de Grenoble)
- Fabrice LEFÈVRE (Avignon Université)
- Eleni METHENITI (UT3-IRIT)
- Véronique MORICEAU (IRIT Université Toulouse 3)
- Leila MOUDJARI (IRIT)
- Alexis NASR (Aix-Marseille Université, LIS)
- Jian-Yun NIE (University of Montreal)
- Damien NOUVEL (INaLCO)
- Aurélie NÉVÉOL (Université Paris-Saclay, CNRS, LISN)
- Magalie OCHS (Aix-Marseille Université, LIS)
- Yannick PARMENTIER (LORIA – Université de Lorraine)
- Patrick PAROUBEK (Université Paris Saclay – CNRS)
- Thierry POIBEAU (LaTTiCe-CNRS)
- Laurent PRÉVOT (Aix Marseille Université, CNRS, Laboratoire Parole et Langage UMR 7309)
- Carlos RAMISCH (Aix Marseille University)
- Sophie ROSSET (Université Paris-Saclay, CNRS, LISN)
- Mickael ROUVIER (Université d’Avignon, LIA)
- Benoît SAGOT (INRIA)
- Eve SAUVAGE (LISN)
- Emmanuel SCHANG (LLL, Univ. Orléans, CNRS)
- Didier SCHWAB (Univ. Grenoble Alpes)

- Djamé SEDDAH (Alpage/Université Paris la Sorbonne)
- Vincent SEGONNE (Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France)
- Gilles SERASSET (LIG – Université Grenoble Alpes)
- Christophe SERVAN (Qwant Research)
- Pascale SÉBILLOT (Université de Rennes, CNRS, Inria / IRISA)
- Ludovic TANGUY (CLLE-ERSS)
- Xavier TANNIER (Sorbonne Université, INSERM, LIMICS)
- Nadi TOMEH (Université Paris 13, Sorbonne Paris Cité)
- Julien TOURILLE (EDF Lab Paris Saclay, SEQUOIA)
- Julien VELCIN (ERIC Lyon 2, EA 3083, Université de Lyon)
- Antoine VENANT (University of Montreal)
- Serena VILLATA (CNRS – Laboratoire d’Informatique, Signaux et Systèmes de Sophia-Antipolis)
- Guillaume WISNIEWSKI (LLF – Université de Paris)
- François YVON (CNRS)
- Gaël DE CHALENDAR (CEA LIST)

Table des matières

« De nos jours, ce sont les résultats qui comptent » : création et étude diachronique d'un corpus de revendications issues d'articles de TAL	1
<i>Clementine Bleuze, Fanny Ducel, Maxime Amblard, Karën Fort</i>	
ALF : Un jeu de données d'analogies françaises à grain fin pour l'évaluation de la connaissance lexicale des grands modèles de langue	22
<i>Alexander Petrov, Antoine Venant, François Lareau, Yves Lepage, Philippe Langlais</i>	
Adaptation des connaissances médicales pour les grands modèles de langue : Stratégies et analyse comparative	50
<i>Ikram Belmadani, Benoit Favre, Richard Dufour, Frédéric Béchet, Carlos Ramisch</i>	
Alignement bi-textuel adaptatif basé sur des plongements multilingues	73
<i>Olivier Kraif</i>	
Alignements divisifs de textes parallèles : données, algorithme et évaluation	84
<i>Joanna Radola, François Yvon</i>	
Alignements entre attention et sémantique dans des modèles de langues pré-entraînés	100
<i>Frédéric Charpentier, Jairo Cugliari Duhalde, Adrien Guille</i>	
Améliorer la Traduction Neuronale par Exemple avec des Données Monolingues	117
<i>Maxime Bouthors, Josep Crego, François Yvon</i>	
Analyse de la continuité référentielle dans le corpus d'écrits scolaires français et italien	134
Scolinter	
<i>Martina Barletta, Claude Ponton</i>	
Augmentation des données par LLM pour améliorer la détection automatique des erreurs de coordination	154
<i>Chunxiao Yan, Iris Eshkol-Taravella, Sarah De Vogué, Marianne Desmets</i>	
Connaissances factuelles dans les modèles de langue : robustesse et anomalies face à des variations simples du contexte temporel	167
<i>Hichem Ammar Khodja, Frédéric Béchet, Quentin Brabant, Alexis Nasr, Gwénolé Lecorvé</i>	
Corpus multilingue annoté pour l'étude sémantique des expressions quantifiantes - Problèmes de segmentation du coréen et du japonais	196
<i>Raoul Blin, Jinnam Choi</i>	
Détecter des comportements associés aux troubles alimentaires par l'analyse automatique des publications textuelles en ligne	206
<i>Yves Ferstler, Catherine Lavoie, Marie-Jean Meurs</i>	
Détection de métaphores dans les documents médicaux	218
<i>Coralie Pottiez, Thierry Hamon, Natalia Grabar</i>	
Détection des contaminations de LLM par extraction de données : une revue de littérature pratique	233
<i>Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, Sophie Rosset</i>	

Détection des omissions dans les résumés médicaux générés par les grands modèles de langue	252
<i>Achir Oukelmoun, Nasredine Semmar, Gaël de Chalendar, Clément Cormi, Mariame Oukelmoun, Eric Vibert, Marc-Antoine Allard</i>	
Détection et évaluation de la communication toxique pour la relation client par des LLMs	268
<i>Guillaume De Murcia, Ludovic Meineri, Laurent Gillard, Thomas Gouritin, Samy Lastmann</i>	
ELITEC : un corpus de conversations en microposts français annoté pour le liage d'entités Wikidata	284
<i>Vivien Leonard, Béatrice Markhoff, Jean-Yves Antoine</i>	
Embeddings, topic models, LLM : un air de famille	295
<i>Ludovic Tanguy, Cécile Fabre, Nabil Hathout, Lydia-Mai Ho-Dac</i>	
Estimation de l'inclusion entre tâches par projection spectrale de vecteurs de tâches	313
<i>Loïc Fosse, Benoît Favre, Frédéric Béchet, Géraldine Damnati, Gwénoél Lecorvé</i>	
Étude comparative de réponses humaines et de grands modèles de langue à des QCM en pharmacie	331
<i>Ricardo Rodriguez, Stéphane Huet, Benoît Favre, Mickael Rouvier</i>	
Étude critique du corpus CNN/DailyMail pour le résumé automatique	348
<i>Fanny Bachev, Christophe Rodrigues, Aurélien Bossard</i>	
Évaluer la capacité des transformeurs à distinguer les significations compositionnelles et idiomatiques d'une même expression	360
<i>Nina Nusbaumer, Guillaume Wisniewski, Benoît Crabbé</i>	
Exploration de la modalité en français parlé et écrit	376
<i>Anna Colli, Delphine Battistelli</i>	
Exploration de la séparation en langues dans les modèles de traitement de la parole auto-supervisés multilingues préentraînés avec des données écologiques	390
<i>William N. Havard, Shrita Hassamal, Muhsina Alleesaib, Guilhem Florigny, Guillaume Fon Sing, Anne Abeillé, Benjamin Lecouteux, Emmanuel Schang</i>	
Identification de mesures d'évaluation fiables pour la révision de textes scientifiques	404
<i>Léane Jourdan, Florian Boudin, Nicolas Hernandez, Richard Dufour</i>	
Intégration des relations inter-référents dans l'annotation de la coréférence : modèle et application	436
<i>Antoine Boiteau, Yann Mathet, Antoine Widlöcher</i>	
L'Impact de la complexité textuelle sur le comportement de lecture : une analyse oculométrique et de la surprise des textes français	450
<i>Oksana Ivchenko, Natalia Grabar</i>	
La confiance de Mistral-7B est-elle justifiée ? Une évaluation en auto-estimation pour les questions biomédicales	467
<i>Laura Zanella, Ambroise Baril</i>	

Latrumplang, instrument de destruction de la pensée : analyse de l'impact de la censure trumpiste sur la recherche en santé mentale	477
<i>Vincent P. Martin, Karën Fort, Jean-Arthur Micoulaud-Franchi</i>	
Le rôle du contexte dans la classification séquentielle de phrases pour les documents longs	488
<i>Anas Belfathi, Nicolas Hernandez, Laura Monceaux, Richard Dufour</i>	
MOSAIC : Mélange d'experts pour la détection de textes artificiels	502
<i>Mathieu Dubois, Pablo Piantanida, François Yvon</i>	
Mesurer les inégalités de genre en ligne avec le genre grammatical : Une étude du subreddit r/france	526
<i>Marie Flesch, Heather Burnett</i>	
Modèles auto-supervisés de traitement de la parole pour le Créole Haitien	542
<i>William N. Havard, Renauld Govain, Benjamin Lecouteux, Emmanuel Schang</i>	
Modélisation de la lisibilité en français pour les personnes en situation d'illettrisme	555
<i>Wafa Aissa, Thibault Bañeras-Roux, Elodie Vanzeveren, Lingyun Gao, Alice Pintard, Rodrigo Wilkens, Thomas François</i>	
Pensez : Moins de données, meilleur raisonnement - Repenser les LLM français	573
<i>Huy Hoang Ha</i>	
Peut-on retrouver votre âge à partir de la transcription de votre parole ?	599
<i>Vanessa Gaudray Bouju, Menel Mahamdi, Iris Eshkol-Taravella, Angèle Barbedette</i>	
Plongement des constituants pour la représentation sémantique des phrases	614
<i>Eve Sauvage, Iskandar Boucharenc, Thomas Gerald, Julien Tourille, Sabrina Campano, Cyril Grouin, Sophie Rosset</i>	
Projeter pour mieux fusionner : une histoire de bandit et de lit	629
<i>Olivier Ferret</i>	
QUARTZ : Approche abstraite non supervisée par question-réponse pour le résumé de dialogue orienté tâche	642
<i>Mohamed Imed Eddine Ghebriout, Gaël Guibon, Ivan Lerner, Emmanuel Vincent</i>	
Raffinage des représentations des tokens dans les modèles de langue pré-entraînés avec l'apprentissage contrastif : une étude entre modèles et entre langues	666
<i>Anna Mosolova, Marie Candito, Carlos Ramisch</i>	
Repousser les limites des benchmarks actuels pour une évaluation réaliste des LLMs en migration de code	682
<i>Samuel Mallet, Joe El Houry, Előd Egyed-Zsigmond</i>	
Supervision faible pour la classification des relations discursives	697
<i>Khalil Maachou, Chloé Braud, Philippe Muller</i>	
Syntaxe en dépendance avec les grammaires catégorielles abstraites : une application à la théorie sens-texte	715
<i>Marie Cousin</i>	

Systèmes d'écriture et qualité des données : l'affinage de modèles de translittération dans un contexte de faibles ressources	729
<i>Emmett Strickland, Ilaine Wang, Damien Nouvel, Bénédicte Diot-Parvaz Ahmad</i>	
Traitement automatique des évènements médiatiques : Détection, classification, segmentation et recherche sémantique	741
<i>Abdelkrim Beloued</i>	
Une revue sur les hallucinations des LLM	756
<i>Eleni Metheniti, Swarnadeep Bhar, Nicholas Asher</i>	
Vers l'entraînement de modèles de reconnaissance automatique de la parole auto-supervisés équitables sans étiquettes démographiques	780
<i>Laura Alonzo-Canul, Benjamin Lecouteux, François Portet</i>	
ding-01 :ARG0 Un corpus AMR pour le français parlé spontané	791
<i>Jeongwoo Kang, Maria Boritchev, Maximin Coavoux</i>	
π-YALLI : un nouveau corpus pour des modèles de langue nahuatl / Yankuik nawat-lahtolkorpus pampa tlahtolmachiotl	802
<i>Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Martha Lorena Avendaño Garrido, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Graham Ranger, Carlos-Emiliano González-Gallardo, Elvys Linhares-Pontes, Patricia Velázquez-Morales, Luis-Gil Moreno-Jiménez</i>	