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 2 : traductions d'articles publiés

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

©2025 ATALA & ARIA

Avec le soutien de

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 précinpale 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 (Universtité Paris 7 / INRIA)
— Vincent CLAVEAU (Irisa)
— Caio CORRO (Université Paris-Saclay)
— Benoît CRABBÉ (Paris 7 et INRIA)
— Iris ESHKOL-TARAVELLA (University of Orléans)
— Karën 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 DUCEL (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 de 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 (Universite Paris Sacle – 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

« Les femmes ne font pas de crise cardiaque ! » Étude des biais de genre dans les cas cliniques synthétiques en français **1**
Fanny Ducel, Nicolas Hiebel, Olivier Ferret, Karën Fort, Aurélie Névéol

ACL-rlg : Un dataset pour la génération de listes de lecture **2**
Julien Aubert-Béduchaud, Florian Boudin, Béatrice Daille, Richard Dufour

AdminSet and AdminBERT : un jeu de données et un modèle de langue pré-entraîné pour explorer le dédale non structuré des données administratives françaises **3**
Thomas Sebbag, Solen Quiniou, Nicolas Stucky, Emmanuel Morin

Anti-surprise : Une métrique complémentaire pour évaluer l'apprentissage lexical des (grands) modèles de langue **5**
Nazanin Shafiabadi, Guillaume Wisniewski

Apprentissage par renforcement pour l'alignement des agents LLMs avec des environnements interactifs : quantification et réduction du surapprentissage aux prompts **6**
Mohamed Salim Aissi, Clement Romac, Thomas Carta, Sylvain Lamprier, Pierre-Yves Oudeyer, Olivier Sigaud, Laure Soulier, Nicolas Thome

Attention Chaînée et Causale pour un Suivi Efficace des Entités **8**
Erwan Fagnou, Paul Caillon, Blaise Delattre, Alexandre Allauzen

Atténuer l'impact de la qualité des références sur l'évaluation des systèmes de résumé grâce aux métriques sans référence **9**
Théo Gigant, Camille Guinaudeau, Marc Decombas, Frédéric Dufaux

Combler les lacunes de Wikipédia : tirer parti de la génération de texte pour améliorer la couverture encyclopédique des groupes sous-représentés **10**
Simon Mille, Massimiliano Pronesti, Craig Thomson, Michela Lorandi, Sophie Fitzpatrick, Rudali Huidrom, Mohammed Sabry, Amy O'Riordan, Anya Belz

EmoDynamiX : Prédiction de stratégies de dialogue pour le support émotionnel via la modélisation de mélange d'émotions et de la dynamique du discours **11**
Chenwei Wan, Matthieu Labeau, Chloé Clavel

Évaluation de la confidentialité des textes cliniques synthétiques générés par des modèles de langue **13**
Foucauld Estignard, Sahar Ghannay, Julien Girard-Satabin, Nicolas Hiebel, Aurélie Névéol

Évaluation des LLMs pour l'Attribution de Citations dans les Textes Littéraires : une Étude de LLaMa3 **14**
Gaspard Michel, Elena V. Epure, Romain Hennequin, Christophe Cerisara

Extraction de mots-clés à partir d'articles scientifiques : comparaison entre modèles traditionnels et modèles de langue **15**
Motasem Alrahabi, Nacef Ben Mansour, Hamed Rahimi

Faut-il éliminer toutes les hallucinations dans un résumé abstractif pour le domaine juridique ? **28**

Nihed Bendahman, Karen Pinel-Sauvagnat, Gilles Hubert, Mokhtar Boumedyen Billami

GeNRe : un système de neutralisation automatique du genre exploitant les noms collectifs **30**
Enzo Doyen, Amalia Todirascu

Graphes, NER et LLMs pour la classification non supervisée de documents **31**
Imed Keraghel, Mohamed Nadif

HISTOIRESMORALES : Un jeu de données français pour évaluer l'alignement moral des modèles de langage **32**
Thibaud Leteno, Irina Proskurina, Antoine Gourru, Julien Velcin, Charlotte Laclau, Guillaume Metzler, Christophe Gravier

Incorporation de Traits de Personnalité dans les Agents Conversationnels à base de GML : Étude de Cas de l'Assistance Client en Français **33**
Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, Fabrice Lefèvre

Inférence en langue naturelle appliquée au recrutement de patients pour les essais cliniques : le point de vue du patient **34**
Mathilde Aguiar, Pierre Zweigenbaum, Nona Naderi

La structure du contenu textuel a-t-elle un impact sur les modèles linguistiques pour le résumé automatique ? **36**
Eve Sauvage, Sabrina Campano, Lydia Ould-Ouali, Cyril Grouin

Lost In Variation : extraction non-supervisée de motifs lexico-syntaxiques dans des textes en moyen arabe **37**
Julien Bezançon, Rimane Karam, Gaël Lejeune

NuNER : Pré-entraînement d'un encodeur pour la reconnaissance d'entités nommées avec des données annotées automatiquement **51**
Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoît Crabbé, Étienne Bernard

PatientDx : Fusion des grands modèles de langue pour la protection de la confidentialité des données dans le domaine de la santé **53**
Jose G. Moreno, Jesús Lovón-Melgarejo, M'Rick Robin-Charlet, Christine-Damase-Michel, Lynda Tamine

Représenter le style au-delà des thématiques : une étude d'impact sur la dispersion vectorielle de différents modèles de langage **55**
Benjamin Icard, Evangelia Zve, Lila Sainero, Alice Breton, Jean-Gabriel Ganascia

SCOPE : un cadre d'entrainement auto-supervisé pour améliorer la fidélité dans la génération conditionnelle de texte **57**
Song Duong, Florian Le Bronnec, Alexandre Allauzen, Vincent Guigue, Alberto Lumbreras, Laure Soulier, Patrick Gallinari

SELEXINI - un grand corpus français, divers et parsé automatiquement **58**
Manon Scholivet, Agata Savary, Louis Estève, Marie Candito, Carlos Ramisch

Sondage des Modèles de Langue sur leur Source de Connaissance **59**
Zineddine Tighidet, Andrea Mogini, Jiali Mei, Patrick Gallinari, Benjamin Piwowarski

Vers les Sens et Au-delà : Induire des Concepts Sémantiques Avec des Modèles de Langue Contextuels **61**
Bastien Liétard, Pascal Denis, Mikaela Keller