

JEP - TALN RECITAL TOULOUSE 2024

*35èmes Journées d'Études sur la Parole (JEP)
31ème Conférence sur le Traitement Automatique des Langues
Naturelles (TALN)
26ème Rencontre des Étudiants Chercheurs en Informatique pour le
Traitement Automatique des Langues (RECITAL)*
<https://jep-taln2024.sciencesconf.org>

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position

Mathieu BALAGUER, Nihed BENDAHDAN, Lydia-Mai HO-DAC, Julie MAUCLAIR, Jose G MORENO,
Julien PINQUIER (Éds.)

Toulouse, France, 8 au 12 juillet 2024

Avec le soutien de



Préface

Organisée conjointement par les équipes de recherche IRIS, MELODI et SAMoVA de l’Institut de Recherche en Informatique de Toulouse (IRIT UMR 5505), l’équipe PLC du laboratoire Cognition, Langues, Langage, Ergonomie (CLLE UMR 5263) et l’axe neurocognition langagière, linguistique et phonétique cliniques du laboratoire de NeuroPsychoLinguistique (LNPL URI EA 4156), sous l’égide de l’Association Francophone de la Communication Parlée (AFCP) et l’Association pour le Traitement Automatique des Langues (ATALA), la conférence JEP-TALN-RECITAL 2024 regroupe :

- les 35^{ème} Journées d’Études sur la Parole (JEP),
- la 31^{ème} Conférence sur le Traitement Automatique des Langues Naturelles (TALN),
- la 26^{ème} Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL).

Les conférences TALN et JEP sont un rendez-vous qui offre le plus important forum d’échange francophone aux acteurs universitaires et industriels des technologies de la langue et la parole. Pour cette édition, nous avons plus de 200 inscrits dont une grande partie des étudiants qui construisent le futur de la recherche francophone et assurent le relais de son développement.

En tant que conférenciers invités, nous aurons Véronique HOSTE de l’Université de Ghent, Laurent BESACIER de Naver Labs Europe et Catia CUCCHIARINI de l’Université de Radboud. Ces trois conférenciers qui représentent un large spectre de thématiques entre le texte et la parole vont aborder les dernières avancées de leurs domaines d’expertise.

Cette édition permet aussi de célébrer les 30 ans de TALN. À cette occasion, nous avons dédié une session spéciale dans le programme. La session a comme objectif de rappeler l’historique de la conférence avec l’intervention des participants qui ont participé à sa pérennité afin de mieux transmettre les enjeux de ce rassemblement à la communauté scientifique du traitement automatique des langues naturelles.

En termes des soumissions, pour TALN, 66 articles pour la conférence principale ont été soumis, dont respectivement 18 ont été acceptés pour une présentation orale et 30 pour une présentation sous forme de posters. Également, nous avons reçu 13 résumés des articles publiés lors de conférences internationales qui ont été acceptés pour une présentation en format poster. En ce qui concerne RECITAL, 11 articles ont été soumis dont 7 ont été acceptés. L’ensemble des soumissions acceptées seront présentées sous forme de posters et 3 d’entre elles donneront lieu à une présentation orale. Pour les JEP, 64 articles ont été soumis et 62 ont été acceptés (17 sous forme de présentation orale et 45 sous format poster). L’alternance de sessions communes entre TALN, JEP et RECITAL et de sessions plus spécifiques devraient permettre de susciter des échanges fructueux. En complément de la conférence principale, se tiennent les ateliers “Parole Spontanée”, “Défi Fouille de Texte” (DEFT), “Jurisprudence Prédictive” (JP’24), “Evaluation des modèles génératifs” (EvalLLM) et l’activité HackaTAL 2024. 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, l'AFCP et le CPRS qui nous ont accompagnés dans les différentes étapes de l'organisation.

Jose G Moreno
Président de TALN

Lydia-Mai Ho-Dac
Nihed Bendahman
Présidentes de RECITAL

Julie Mauclair
Présidente de JEP

Comités

Comité de Programme

- Rachel Bawden, Inria
- Leonor Becerra-Bonache, Laboratoire d'Informatique et Systèmes
- Delphine Bernhard, LiLPa, Université de Strasbourg
- Nathalie Camelin, LIUM — Université du Maine
- Marie Candito, Université Paris 7 / INRIA
- Vincent Claveau, Irisa
- Géraldine Damnati, Orange Labs
- Iris Eshkol-Taravella, University of Orléans
- Benoit Favre, Aix-Marseille Université
- Natalia Grabar, STL CNRS Université Lille 3
- Thierry Hamon, France
- Lydia-Mai Ho-Dac, CLLE
- Philippe Langlais, Canada
- Jose G Moreno, IRIT – Université Paul Sabatier
- Emmanuel Morin, Université de Nantes, LS2N
- Vincent Segonne, Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France
- Christophe Servan, Qwant Research
- Anne Vilnat, LIMSI-CNRS

Comité de Relecture

- Maxime Amblard, Université de Lorraine
- Jean-Yves Antoine, Université François Rabelais de Tours
- Lauriane Aufrant, Inria
- Frederic Bechet, Aix Marseille Université - LIF
- Patrice Bellot, Aix-Marseille Université - CNRS (LIS)
- Asma Ben Abacha, Microsoft Health AI
- Timothée Bernard, Université Paris Cité
- Romaric Besançon, CEA LIST
- Philippe Blache, LPL, AMU
- Chloé Braud, IRIT - CNRS
- Remi Cardon, CENTAL, IL&C, Université Catholique de Louvain
- Maximin Coavoux, CNRS, Université Grenoble Alpes
- Matthieu Constant, Université de Lorraine, ATILF, CNRS
- Caio Corro, Université Paris-Saclay
- Benoît Crabbé, Paris 7 et INRIA
- Béatrice Daille, Laboratoire d'Informatique Nantes Atlantique (LINA)
- Gaël de Chalendar, CEA LIST
- Gaël Dias, Normandie University
- Taoufiq Dkaki, IRIT, Institut de Recherche en Informatique de Toulouse
- Benamara Farah, Univ. Paul Sabatier, Toulouse and IPAL, Singapore
- Olivier Ferret, CEA List
- Karën Fort, Sorbonne Université
- Amel Fraisse, Université de Lille
- Thomas Francois, Université catholique de Louvain
- Sahar Ghannay, LISN lab
- Cyril Grouin, LISN

- Gaël Guibon, Université de Lorraine - LORIA
- Nabil Hathout, CNRS
- Nicolas Hernandez, Nantes Université - LS2N CNRS UMR 6004
- Gilles Hubert, IRIT
- Luce Lefeuvre, DTIPG, SNCF
- Fabio Martínez Carrillo, Bivl2ab- Biomedical Imaging, vision and learning laboratory. Universidad Industrial de Santander
- Véronique Moriceau, IRIT Université Toulouse 3
- Philippe Muller, IRIT, Toulouse University
- Alexis Nasr, LIS
- Aurélie Névéol, Université Paris-Saclay, CNRS, LISN
- Jian-Yun Nie, University de Montreal
- Damien Nouvel, INALCO
- Yannick Parmentier, LORIA - Université de Lorraine
- Patrick Paroubek, Université Paris Saclay - CNRS
- Benjamin Piwowarski, CNRS / ISIR, Sorbonne Université
- Thierry Poibeau, LaTTiCe-CNRS
- Solen Quiniou, LS2N - Nantes Université
- Benoît Sagot, INRIA
- Djamé Seddah, Alpage/Université Paris la Sorbonne
- Nasredine Semmar, CEA
- Ludovic Tanguy, CLLE-ERSS
- Xavier Tannier, Sorbonne Université, INSERM, LIMICS
- Julien Tourille, CEA, LIST
- Guillaume Wisniewski, LLF - Université de Paris
- François Yvon, CNRS
- Pierre Zweigenbaum, Université Paris-Saclay, CNRS, LISN

Table des matières

I	Articles présentés oralement	1
	À propos des difficultés de traduire automatiquement de longs documents	2
	<i>Ziqian Peng, Rachel Bawden, François Yvon</i>	
	Approches cascade et de bout-en-bout pour la traduction automatique de la parole en pictogrammes	22
	<i>Cécile Macaire, Chloé Dion, Didier Schwab, Benjamin Lecouteux, Emmanuelle Esperança-Rodier</i>	
	Au-delà de la performance des modèles : la prédiction de liens peut-elle enrichir des graphes lexico-sémantiques du français ?	36
	<i>Hee-Soo Choi, Priyansh Trivedi, Mathieu Constant, Karën Fort, Bruno Guillaume</i>	
	CQuAE : Un nouveau corpus de question-réponse pour l’enseignement	50
	<i>Thomas Gerald, Louis Tamames, Sofiane Ettayeb, Patrick Paroubek, Anne Vilnat</i>	
	Évaluation automatique des biais de genre dans des modèles de langue auto-régressifs	64
	<i>Fanny Ducl, Aurélie Névéol, Karën Fort</i>	
	Évaluation de la Similarité Textuelle : Entre Sémantique et Surface dans les Représentations Neuronales	85
	<i>Julie Tytgat, Guillaume Wisniewski, Adrien Betrancourt</i>	
	Extraction des arguments d’événements à partir de peu d’exemples par méta-apprentissage	97
	<i>Aboubacar Tuo, Romaric Besançon, Olivier Ferret, Julien Tourille</i>	
	Les petits modèles sont bons : une étude empirique de classification dans un contexte zero-shot	113
	<i>Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, Sophie Rosset</i>	
	Les représentations contextuelles stéréotypées dans les modèles de langue français : mieux les identifier pour ne pas les reproduire	130
	<i>Léandre Adam-Cuvillier, Pierre-Jean Larpin, Antoine Simoulin</i>	
	Méta-apprentissage pour l’analyse AMR translingue	144
	<i>Jeongwoo Kang, Maximin Coavoux, Cédric Lopez, Didier Schwab</i>	
	Recherche de relation à partir d’un seul exemple fondée sur un modèle N-way K-shot : une histoire de distracteurs	157
	<i>Hugo Thomas, Guillaume Gravier, Pascale Sébillot</i>	
	Reconnaissance d’entités cliniques en few-shot en trois langues	169
	<i>Marco Naguib, Aurélie Névéol, Xavier Tannier</i>	
	Réduction des répétitions dans la Traduction Automatique Neuronale	198
	<i>Marko Avila, Anna Rebollo, Josep Crego</i>	
	Régression logistique parcimonieuse pour l’extraction automatique de règles de grammaire	211
	<i>Santiago Herrera, Caio Corro, Sylvain Kahane</i>	

SEC : contexte émotionnel phrastique intégré pour la reconnaissance émotionnelle efficiente dans la conversation	219
<i>Barbara Gendron, Gaël Guibon</i>	
Une approche par graphe pour l'analyse syntaxique en dépendances de bout en bout de la parole	234
<i>Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, Jérôme Goulian</i>	
Vers la traduction automatique des néologismes scientifiques	245
<i>Paul Lerner, François Yvon</i>	
WikiFactDiff : Un Grand jeu de données Réaliste et Temporellement Adaptable pour la Mise à Jour Atomique des Connaissances Factuelles dans les Modèles de Langue Causaux	262
<i>Hichem Ammar Khodja, Frédéric Béchet, Quentin Brabant, Alexis Nasr, Gwénolé Lecrové</i>	
II Articles présentés en session poster	282
Adaptation des modèles de langue à des domaines de spécialité par un masquage sélectif fondé sur le genre et les caractéristiques thématiques	283
<i>Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, Richard Dufour</i>	
Améliorer la traduction au niveau du document grâce au sur-échantillage négatif et au masquage ciblé	295
<i>Gaëtan Caillaud, Mariam Nakhlé, Jingshu Liu, Raheel Qader</i>	
Améliorer les modèles de langue pour l'analyse des émotions : perspectives venant des sciences cognitives	307
<i>Constant Bonard, Gustave Cortal</i>	
Analyse de la perception de l'offre INTERCITÉS de jour : Classification multi-étiquettes des émotions dans les tweets	323
<i>Chang Liu, Hélène Flamein, Luce Lefevre, Fanny Hanen</i>	
Approche multitâche pour l'amélioration de la fiabilité des systèmes de résumé automatique de conversation	338
<i>Eunice Akani, Benoit Favre, Frederic Bechet, Romain Gemignani</i>	
Auto-correction et oracle dynamique : certains effets n'apparaissent qu'à taille réduite	352
<i>Fang Zhao, Timothée Bernard</i>	
Construction d'une mesure de similarité thématique non supervisée pour les conversations	362
<i>Amandine Decker, Maxime Amblard</i>	
De nouvelles méthodes pour l'exploration de l'interface syntaxe-prosodie : un treebank intonosyntaxique et un système de synthèse pour le pidgin nigérian	376
<i>Emmett Strickland, Anne Lacheret-Dujour, Marc Evrard, Sylvain Kahane, Dana Aubakirova, Dorin Doncenco, Diego Torres, Perrine Quennehen, Bruno Guillaume</i>	
Étude des facteurs de complexité des modèles de langage dans une tâche de compréhension de lecture à l'aide d'une expérience contrôlée sémantiquement	384

Elie Antoine, Frederic Bechet, Géraldine Damnati, Philippe Langlais

- Évaluation de l'apport des chaînes de coréférences pour le liage d'entités** 397
Léo Labat, Lauriane Aufrant
- Extension d'AZee avec des règles de production concernant les gestes non-manuels pour la langue des signes française** 410
Camille Challant, Michael Filhol
- Extraction d'entités nommées décrivant des chaînes de traitement bioinformatiques dans des articles scientifiques en anglais** 422
Clémence Sebe, Sarah Cohen-Boulakia, Olivier Ferret, Aurélie Névéol
- Génération contrôlée de cas cliniques en français à partir de données médicales structurées** 435
Hugo Boulanger, Nicolas Hiebel, Olivier Ferret, Karën Fort, Aurélie Névéol
- L'impact de genre sur la prédiction de la lisibilité du texte en FLE** 449
Lingyun Gao, Rodrigo Wilkens, Thomas François
- LLM-Generated Contexts to Practice Specialised Vocabulary : Corpus Presentation and Comparison** 472
Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin, Thomas François
- La reconnaissance automatique des relations de cohérence RST en français.** 499
Martial Pastor, Erik Bran Marino, Nelleke Oostdijk
- MEETING : A corpus of French meeting-style conversations** 508
Julie Hunter, Hiroyoshi Yamasaki, Océane Granier, Jérôme Louradour, Roxane Bertrand, Kate Thompson, Laurent Prévot
- MODEL : Large Language Models for Spontaneous French Dialogue** 530
Jérôme Louradour, Julie Hunter, Ismaïl Harrando, Guokan Shang, Virgile Rennard, Jean-Pierre Lorré
- Modéliser la facilité d'écoute en FLE : vaut-il mieux lire la transcription ou écouter le signal vocal ?** 549
Minami Ozawa, Rodrigo Wilkens, Kaori Sugiyama, Thomas François
- Optimisation des performances d'un système de reconnaissance automatique de la parole pour les commentaires sportifs : fine-tuning de Whisper** 567
Camille Lavigne, Alex Stasica, Anna Kupsc
- Optimiser le choix des exemples pour la traduction automatique augmentée par des mémoires de traduction** 582
Maxime Bouthors, Josep Crego, François Yvon
- ParaPLUIE - une mesure automatique d'évaluation de la qualité sémantique des systèmes de paraphrases** 605
Quentin Lemesle, Jonathan Chevelu, Damien Lolive, Arnaud Delhay-Lorrain, Philippe Martin
- Prédiction de la complexité lexicale : Une étude comparative entre ChatGPT et un modèle dédié à cette tâche.** 617

Abdelhak Kelious, Mathieu Constant, Christophe Coeur

Quel workflow pour les sciences du texte ? 630

Antoine Widlöcher

Repérage et caractérisation automatique des émotions dans des textes : traiter aussi leurs modes d'expression indirects 650

Aline Etienne, Delphine Battistelli, Gwénoél Lecorvé

TCFLE-8 : un corpus de productions écrites d'apprenants de français langue étrangère et son application à la correction automatisée de textes 677

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, Thomas François

Technologies de la parole et données de terrain : le cas du créole haïtien 686

William N. Havard, Renauld Govain, Daphne Gonçalves Teixeira, Benjamin Lecouteux, Emmanuel Schang

Utiliser l'explicabilité des modèles pour mettre en évidence les expressions genrées dans la parole 695

François Buet, Camille Guinaudeau, Cyril Grouin, Sahar Ghannay, Shin'Ichi Satoh

Vers une pédagogie inclusive : une classification multimodale des illustrations de manuels scolaires pour des environnements d'apprentissage adaptés 708

Saumya Yadav, Élise Lincker, Caroline Huron, Stéphanie Martin, Camille Guinaudeau, Shin'Ichi Satoh, Jainendra Shukla

astroECR : enrichissement d'un corpus astrophysique en entités nommées, coréférences et relations sémantiques 720

Atilla Kaan Alkan, Felix Grezes, Cyril Grouin, Fabian Schüssler, Pierre Zweigenbaum

Première partie

Articles présentés oralement