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 2 : traductions d'articles publiés

Mathieu BALAGUER, Nihed BENDAHMAN, 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 35ème Journées d'Études sur la Parole (JEP),
— la 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN),
— la 26ème 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	Lydia-Mai Ho-Dac Nihed Bendahman	Julie Mauclair
Président de TALN	Présidentes de RECITAL	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, Universtité 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 Franois 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, Intitut 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, Universite Paris Sacle - 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

Apport de la structure de tours à l'identification automatique de genre textuel : un corpus annoté de sites web de tourisme en français **1**
Remi Cardon, Trang Tran Hanh Pham, Julien Zakhia Doueihi, Thomas François

Caractérisation de la ville du futur dans un corpus de science-fiction **2**
Sami Guembour, Chuanming Dong, Catherine Dominguès

ChiCA : un corpus de conversations face-à-face vs. Zoom entre enfants et parents **4**
Dhia Elhak Goumri, Abhishek Agrawal, Mitja Nikolaus, Hong Duc Thang Vu, Kübra Bodur, Elias Semmar, Cassandre Armand, Chiara Mazzocconi, Shreejata Gupta, Laurent Prévot, Benoit Favre, Leonor Becerra-Bonache, Abdellah Fourtassi

Évaluer les modèles de langue pré-entraînés avec des propriétés de hiérarchie **6**
Jesus Lovon-Melgarejo, Jose G Moreno, Romaric Besançon, Olivier Ferret, Lynda Tamine

Exploration d'approches hybrides pour la lisibilité : expériences sur la complémentarité entre les traits linguistiques et les transformers **8**
Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, Thomas François

Jargon : Une suite de modèles de langues et de référentiels d'évaluation pour les domaines spécialisés du français **9**
Vincent Segonne, Aidan Mannion, Laura Alonzo-Canul, Audibert Alexandre, Xingyu Liu, Cécile Macaire, Adrien Pupier, Yongxin Zhou, Mathilde Aguiar, Felix Herron, Magali Norré, Massih-Reza Amini, Pierrette Bouillon, Iris Eshkol Taravella, Emmanuelle Esparança-Rodier, Thomas François, Lorraine Goeuriot, Jérôme Goulian, Mathieu Lafourcade, Benjamin Lecouteux, François Portet, Fabien Ringeval, Vincent Vandeghinste, Maximin Coavoux, Marco Dinarelli, Didier Schwab

LOCOST : Modèles Espace-État pour le Résumé Abstractif de Documents Longs **11**
Florian Le Bronnec, Song Duong, Alexandre Allauzen, Vincent Guigue, Alberto Lumbreras, Laure Soulier, Patrick Gallinari

La subjectivité dans le journalisme québécois et belge : transfert de connaissance inter-médias et inter-cultures **12**
Louis Escouflaire, Antonin Descampe, Antoine Venant, Cédrick Fairon

Le corpus BrainKT : Etudier l'instanciation du common ground par l'analyse des indices verbaux, gestuels et neurophysiologiques **14**
Eliot Maës, Thierry Legou, Leonor Becerra-Bonache, Philippe Blache

Rééquilibrer la distribution des labels tout en éliminant le temps d'attente inhérent dans l'apprentissage actif multi-label appliqué aux transformers **16**
Maxime Arens, Jose G Moreno, Mohand Boughanem, Lucile Callebert

Sur les limites de l'identification par l'humain de textes générés automatiquement **18**
Nadége Alavoine, Maximin Coavoux, Emmanuelle Esperança-Rodier, Romane Gallienne, Carlos-Emiliano González-Gallardo, Jérôme Goulian, Jose G Moreno, Aurélie Névéol, Didier Schwab, Vincent Segonne, Johanna Simoens

Un corpus multimodal alignant parole, transcription et séquences de pictogrammes dédié à la traduction automatique de la parole vers des pictogrammes **20**

Cécile Macaire, Chloé Dion, Jordan Arrigo, Claire Lemaire, Emmanuelle Esperança-Rodier, Benjamin Lecouteux, Didier Schwab

Une approche zero-shot pour localiser les transferts d'informations en conversation naturelle **22**
Eliot Maës, Hossam Boudraa, Philippe Blache, Leonor Becerra-Bonache