



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
*(CORIA-TALN)*¹

Actes de CORIA-TALN 2023.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles
(TALN),
volume 4 : articles déjà soumis ou acceptés en conférence internationale

Christophe Servan, Anne Vilnat (Éds.)

Paris, France, 5 au 9 juin 2023

1. <https://coria-taln-2023.sciencesconf.org/>

Avec le soutien de



Préface

Organisée conjointement par les laboratoires franciliens sous l'égide de l'Association francophone de Recherche d'Information et Applications (ARIA) et l'Association pour le Traitement Automatique des Langues (ATALA), la conférence CORIA-TALN-RJCRI-RECITAL 2023 regroupe :

- la 18ème Conférence en Recherche d'Information et Applications (CORIA)
 - la 30ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) ;
- ainsi que les deux conférences associées, destinées aux jeunes chercheuses et chercheurs :
- Les 16ème Rencontres Jeunes Chercheurs en RI (RJCRI)
 - la 25ème Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)

La conférence TALN (Traitement Automatique des Langues Naturelles) est un rendez-vous annuel qui offre, depuis 1994, le plus important forum d'échange international francophone aux acteurs universitaires et industriels des technologies de la langue. Cet événement, qui accueille habituellement près de 250 participants, couvre toutes les avancées récentes en matière de communication écrite et parlée et de traitement informatique de la langue notamment la recherche et l'extraction d'information, la fouille de textes, le dialogue homme-machine, la fouille d'opinions, la traduction automatique, les systèmes de questions-réponses, le résumé automatique...

Cette année, ont été soumis 51 articles longs et 12 articles courts pour la conférence principale, dont respectivement 29 ont été acceptés pour une présentation orale (dont 2 prises de position) et 9 pour une présentation sous forme de posters. 19 présentations courtes, sous forme de posters, d'articles déjà publiés lors de conférences internationales complètent le programme de la conférence, ainsi que des démonstrations et des présentations de projets en cours. L'alternance de sessions communes entre TALN, CORIA et RJC et de sessions plus spécifiques devraient permettre de susciter des échanges fructueux.

En complément de la conférence principale, se tiennent les ateliers "Défi Fouille de Texte" (DEFT), "Atelier sur l'analyse et la recherche de textes scientifiques" (ARTS), "Humain ou pas humain ? : les nouveaux défis pour les humains" (hOUPSh) et le tutoriel "Apprentissage Profond pour le TAL français pour les débutants" (TutoriAL). Ces ateliers et tutoriel illustrent à la fois des tendances nouvelles présentes dans la communauté et des activités récurrentes.

Un grand merci à toutes celles et tous ceux qui ont soumis leurs travaux, ainsi qu'aux membres du comité de programme et aux relectrices et relecteurs pour le travail qu'ils ont accompli. Ce sont eux qui font vivre la conférence. Merci au comité d'organisation réparti sur la région parisienne, et aux sponsors qui nous ont permis d'organiser cet événement.

Christophe Servan et Anne Vilnat, co-présidents de TALN

Comités

Comité de programme

Présidents

- Christophe SERVAN
- Anne VILNAT

Membres

- Rachel BAWDEN
- Caroline BRUN
- Marie CANDITO
- Rémi CARDON
- Pascal DENIS
- Yannick ESTEVE
- Benoît FAVRE
- Amel FRAISSE
- Thomas GERALD
- Natalia GRABAR
- Lydia-Mai HO-DAC
- José MORENO
- Vassilina NIKOULINA
- Yannick PARMENTIER
- Sylvain POGODALLA
- Solène QUINIOU
- Didier SCHWAB
- Iris TARAVELLA-ESHKOL

Comité d'organisation

- Marie CANDITO
- Thomas GERALD
- José MORENO
- Benjamin PIWOWARSKI
- Christophe SERVAN
- Laure SOULIER
- Anne VILNAT

Table des matières

Questionner pour expliquer : construction de liens explicites entre documents par la génération automatique de questions	1
<i>Elie Antoine, Hyun Jung Kang, Ismaël Rousseau, Ghislaine Azémard, Frédéric Béchet, Géraldine Damnati</i>	
HATS : Un jeu de données intégrant la perception humaine appliquée à l'évaluation des métriques de transcription de la parole	10
<i>Thibault Bañeras-Roux, Jane Wottawa, Mickael Rowvier, Teva Merlin, Richard Dufour</i>	
Résumé automatique multi-documents guidé par une base de résumés similaires	19
<i>Florian Baud, Alexandre Aussem</i>	
Traduction à base d'exemples du texte vers une représentation hiérarchique de la langue des signes	28
<i>Elise Bertin-Lemée, Annelies Braffort, Camille Challant, Claire Danet, Michael Filhol</i>	
Annotation Linguistique pour l'Évaluation de la Simplification Automatique de Textes	35
<i>Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Patrick Watrin, Thomas François</i>	
Un mot, deux facettes : traces des opinions dans les représentations contextualisées des mots	49
<i>Aina Garí Soler, Matthieu Labeau, Chloe Clavel</i>	
PromptORE - Vers l'Extraction de Relations non-supervisée	58
<i>Pierre-Yves Genest, Pierre-Edouard Portier, Előd Egyed-Zsigmond, Laurent-Walter Goix</i>	
Injection de connaissances temporelles dans la reconnaissance d'entités nommées historiques	65
<i>Carlos-Emiliano González-Gallardo, Emanuela Boros, Edward Giamphy, Ahmed Hamdi, Jose Moreno, Antoine Doucet</i>	
Oui mais... ChatGPT peut-il identifier des entités dans des documents historiques ?	74
<i>Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose Moreno, Antoine Doucet</i>	
De l'interprétabilité des dimensions à l'interprétabilité du vecteur : parcimonie et stabilité	83
<i>Simon Guillot, Thibault Prouteau, Nicolas Dugue</i>	
Effet de l'anthropomorphisme des machines sur le français adressé aux robots : Étude du débit de parole et de la fluence	92
<i>Natalia Kalashnikova, Mathilde Hutin, Ioana Vasilescu, Laurence Devillers</i>	
Détection de la nasalité du locuteur à partir de réseaux de neurones convolutifs et validation par des données aérodynamiques	101
<i>Lila Kim, Cedric Gendrot, Amélie Elmerich, Angélique Amelot, Shinji Maeda</i>	
DrBERT : Un modèle robuste pré-entraîné en français pour les domaines biomédical et clinique	109

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, Pierre-Antoine Gourraud

Classification automatique de données déséquilibrées et bruitées : application aux exercices de manuels scolaires **121**

Elise Lincker, Camille Guinaudeau, Olivier Pons, Jérôme Dupire, Isabelle Barbet, Céline Hudelot, Vincent Mousseau, Caroline Huron

Détection de faux tickets de caisse à l'aide d'entités et de relations basées sur une ontologie de domaine **131**

Beatriz Martínez Tornés, Emanuela Boros, Petra Gomez-Krämer, Antoine Doucet, Jean-Marc Ogier

Jeu de données de tickets de caisse pour la détection de fraude documentaire **140**

Beatriz Martínez Tornés, Théo Taburet, Emanuela Boros, Kais Rouis, Petra Gomez-Krämer, Nicolas Sidere, Antoine Doucet, Vincent Poulain D'andecy

Portabilité linguistique des modèles de langage pré-appris appliqués à la tâche de dialogue humain-machine en français **148**

Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, Fabrice Lefèvre

Détection d'événements à partir de peu d'exemples par seuillage dynamique **159**

Aboubacar Tuo, Romaric Besançon, Olivier Ferret, Julien Tourille

Sélection globale de segments pour la reconnaissance d'entités nommées **169**

Urchade Zaratiana, Niama El Khbir, Pierre Holat, Nadi Tomeh, Thierry Charnois