



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 1 : travaux de recherche originaux – articles longs

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

Étude de méthodes d’augmentation de données pour la reconnaissance d’entités nommées en astrophysique	1
<i>Atilla Kaan Alkan, Cyril Grouin, Pierre Zweigenbaum</i>	
Towards a Robust Detection of Language Model-Generated Text : Is ChatGPT that easy to detect ?	14
<i>Wissam Antoun, Virginie Mouilleron, Benoît Sagot, Djamé Seddah</i>	
Cross-lingual Strategies for Low-resource Language Modeling : A Study on Five Indic Dialects	28
<i>Niyati Bafna, Cristina España-Bonet, Josef Van Genabith, Benoît Sagot, Rachel Bawden</i>	
Pauee : Prédiction des pauses dans la lecture d’un texte	43
<i>Marion Baranes, Karl Hayek, Romain Hennequin, Elena V. Epure</i>	
Reconnaissance de défigements dans des tweets en français par des mesures de similarité sur des alignements textuels	56
<i>Julien Bezançon, Gaël Lejeune</i>	
Tri-apprentissage génératif : génération de données pour de la reconnaissance d’entités nommées semi-supervisé	68
<i>Hugo Boulanger, Thomas Lavergne, Sophie Rosset</i>	
Évaluation d’un générateur automatique de reformulations médicales	80
<i>Ioana Buhnila, Amalia Todirascu</i>	
Étude comparative des plongements lexicaux pour l’extraction d’entités nommées en français	94
<i>Danrun Cao, Nicolat Béchet, Pierre-François Marteau</i>	
”Honey, Tell Me What’s Wrong”, Explicabilité Globale des Modèles de TAL par la Génération Coopérative	105
<i>Antoine Chaffin, Julien Delaunay</i>	
Extraction de relations sémantiques et modèles de langue : pour une relation à double sens	123
<i>Olivier Ferret</i>	
Géométrie de l’auto-attention en classification : quand la géométrie remplace l’attention	137
<i>Loïc Fosse, Duc Hau Nguyen, Pascale Sébillot, Guillaume Gravier</i>	
Un traitement hybride du vague textuel : du système expert VAGO à son clone neuronal	151
<i>Benjamin Icard, Vincent Claveau, Ghislain Ateazing, Paul Egré</i>	
Uniformité de la densité informationnelle : le cas du redoublement du sujet	164
<i>Yiming Liang, Pascal Amsili, Heather Burnett</i>	
Augmentation des modèles de langage français par graphes de connaissances pour la reconnaissance des entités biomédicales	177
<i>Aidan Mannion, Schwab Didier, Lorraine Goeuriot, Thierry Chevalier</i>	

Annotation d'entités cliniques en utilisant les Larges Modèles de Langue	190
<i>Simon Meoni, Théo Ryffel, Eric De La Clergerie</i>	
Classification de tweets en situation d'urgence pour la gestion de crises	204
<i>Romain Meunier, Leila Moudjari, Farah Benamara, Véronique Moriceau, Alda Mari, Patricia Stolf</i>	
Outils de l'occitan : nouvelles ressources et lemmatisation	217
<i>Aleksandra Miletić</i>	
Stratégies d'apprentissage actif pour la reconnaissance d'entités nommées en français	232
<i>Marco Naguib, Aurélie Névéol, Xavier Tannier</i>	
Détecter une erreur dans les phrases coordonnées au sein des rédactions universitaires	248
<i>Laura Noreskal, Iris Eshkol-Taravella, Marianne Desmets</i>	
Production automatique de gloses interlinéaires à travers un modèle probabiliste exploitant des alignements	262
<i>Shu Okabe, François Yvon</i>	
Intégration de connaissances structurées par synthèse de texte spécialisé	275
<i>Guilhem Piat, Ellington Kirby, Julien Tourille, Nasredine Semmar, Alexandre Allauzen, Hassane Essafi</i>	
DACCORD : un jeu de données pour la Détection Automatique d'énoncés Contra-	
Dictoires en français	285
<i>Maximos Skandalis, Richard Moot, Simon Robillard</i>	
Exploitation de plongements de graphes pour l'extraction de relations biomédicales	298
<i>Anfu Tang, Robert Bossy, Louise Deléger, Claire Nédellec, Pierre Zweigenbaum</i>	
Derrière les plongements de relations	311
<i>Hugo Thomas, Guillaume Gravier, Pascale Sébillot</i>	
CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé	323
<i>Rian Touchent, Laurent Romary, Eric De La Clergerie</i>	
Protocole d'annotation multi-label pour une nouvelle approche à la génération de réponse socio-émotionnelle orientée-tâche	335
<i>Lorraine Vanel, Alya Yacoubi, Chloe Clavel</i>	
Exploring Data-Centric Strategies for French Patent Classification : A Baseline and Comparisons	349
<i>You Zuo, Benoît Sagot, Kim Gerdes, Houda Mouzoun, Samir Ghamri Doudane</i>	