



Traitement Automatique des Langues Naturelles
(TALN)¹

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : conférence principale

Pascal Denis, Natalia Grabar, Amel Fraise, Rémi Cardon, Bernard Jacquemin, Eric Kergosien, Antonio Balvet
(Éds.)

Lille, France, 28 juin au 2 juillet 2021

1. <https://talnrecital2021.inria.fr/>

Avec le soutien de

Soutiens institutionnels



Sponsors industriels

Partenaires « Argent »



Partenaires « Bronze »



Préface

Pour sa 28e édition, la conférence TALN s’est tenue pour la première fois de son histoire à Lille, et pour la seconde fois seulement dans la région des Hauts-de-France (après TALN 2009 à Senlis). Comme il en est devenu la tradition, TALN est une nouvelle fois organisée sous l’égide de l’ATALA conjointement avec sa conférence “soeur”, RÉCITAL, dont c’est déjà la 23e édition.

Comme pour leurs éditions 2020, TALN 2021 et RÉCITAL 2021 ont à nouveau dû être “virtualisées” en raison de l’épidémie de Covid-19 qui a paralysé la France, l’Europe, et une bonne partie du monde. Ceci a considérablement compliqué son organisation et a conduit à la suppression de plusieurs événements originellement prévus dont le HackaTAL, la soirée gala, les événements sociaux, les promenades à Lille, la dégustation de la cuisine régionale, etc. Néanmoins, nous avons pu maintenir l’atelier Défi Fouilles de Textes (DEFT 2021), ainsi que non moins de 8 tutoriels différents. Nous remercions les organisateurs vaillants de DEFT et des tutoriels.

En lien avec cette actualité sanitaire, le thème choisi pour l’édition de TALN 2021 est “TAL et santé”. Ce thème se reflète naturellement dans le programme de cette édition, puisqu’elle comprend une conférence invitée de Pierre Zweigenbaum sur le TAL médical, une session dédiée, et le traitement des cas cliniques comme tâche de DEFT 2021. Nous avons par ailleurs été très contents d’accueillir André Martins (professeur associé à l’Instituto Superior Técnico et VP recherche chez Unbabel, à Lisbonne au Portugal), comme second conférencier invité de cette édition.

Ces actes regroupent les articles des conférences TALN et RÉCITAL (volume 1 et 2, respectivement), ceux décrivant les démonstrations (volume 3), ceux issus de l’atelier DEFT 2021 (volume 4). Comme lors de la précédente édition de TALN 2020, un appel spécifique réservé aux résumés d’articles publiés dans des conférences internationales de premier plan fut également organisé. Ces résumés ont été versés dans le volume 1.

Pour TALN, un total de 58 articles a été soumis, soit exactement le même nombre que pour l’édition précédente. Parmi ceux-ci, 45 ont été sélectionnés, soit un taux d’acceptation de 77.6 %, dont 8 comme articles longs et 37 comme articles courts. Pour RÉCITAL, le nombre d’articles soumis fut de 16, en léger recul par rapport aux 22 soumissions de l’an dernier. 13 de ceux-ci ont été sélectionnés, soit un taux d’acceptation de 81.2 %.

Parmi les innovations de cette édition de TALN-RÉCITAL, nous avons rajouté une phase de discussion entre auteur(e)s et relecteurs/relectrices, de manière à enrichir et fluidifier le processus de relecture et, on l’espère, à améliorer la sélection des articles et la plus-value des retours apportée aux auteur(e)s.

Nous sommes extrêmement reconnaissants à toutes les personnes qui ont participé aux différents comités scientifiques de ces conférences, à savoir :

- les responsables de domaine de TALN (voir page [vi](#)) ;
- les relectrices et relecteurs de TALN et RÉCITAL (voir page [vi](#)).

En outre, nous remercions chaleureusement l’ATALA, dont le comité permanent (le CPerm) assure la pérennité des TALN et RÉCITAL. Nous sommes également redevables à l’ensemble des membres du comité d’organisation (en particulier Antonio Balvet et Bernard Jacquemin), ainsi qu’aux personnes qui ont apporté leur soutien administratif et logistique

(en particulier Christine Yvoz) pour leur implication. Merci aussi à Yannick Parmentier qui nous a permis de produire ces actes et d'assurer la diffusion de ceux-ci sur HAL, l'ACL anthology et les archives TALN. Nous remercions aussi Onkar Pandit, Mariana Vargas et Nathalie Vauquier pour leur aide dans la maintenance du site web de la conférence et dans la configuration de la plate-forme `gather.town`.

Enfin, que soient aussi remerciés nos partenaires institutionnels et industriels pour leur soutien financier, en particulier : le CNRS, l'Inria, l'Université de Lille, les laboratoires CRISAL, STL et GERIICO, l'ATALA et l'Afia, la GDLFLF, et les entreprises Schlumberger, ELRA, ERDIL, SINEQUA, ZENDOC.

Les présidentes et présidents de TALN : Pascal Denis et Natalia Grabar ;

Les présidentes et présidents de RÉCITAL : Amel Fraisse et Rémi Cardon.

Comités

Co-Président.e.s TALN

- Pascal Denis, MAGNET, Inria Lille & CRIStAL
- Natalia Grabar, STL, CNRS

Responsables de domaine

- Delphine Bernhard, LiLPA, Strasbourg
- Houda Bouamor, CMU Qatar
- Chloé Braud, IRIT, Toulouse
- Caroline Brun, NaverLabs, Grenoble
- Marie Candito, LLF, Paris
- Caio Corro, LISN, CNRS, Université Paris-Saclay
- Géraldine Damnati, Orange Labs, Lannion
- Maud Erhmann, EPFL, Suisse
- Cécile Fabre, CLLE, Toulouse
- Benoît Favre, TALEP, Marseille
- Thomas François, CENTAL, UCLouvain, Louvain-la-Neuve, Belgique
- Nuria Gala, LPL, Aix
- Philippe Langlais, DIRO, Montréal, Canada
- Philippe Muller, IRIT, Toulouse
- Alexis Nasr, TALEP, Marseille
- Magalie Ochs, LIS, Marseille
- Yannick Parmentier, LORIA, Nancy
- Tim van de Cruys, KUL, Leuven, Belgique
- Guillaume Wisniewski, LLF, Paris

Comité de lecture TALN

- Céline Alec, GREYC, Université de Caen-Normandie
- Alexandre Allauzen, LAMSADE, Université Paris-Dauphine
- Maxime Amblard, LORIA, Université de Lorraine

- Pascal Amsili, LATTICE, ILPGA, Université Sorbonne Nouvelle
- Loïc Barrault, University of Sheffield
- Patrice Bellot, Aix-Marseille Université – CNRS (LIS)
- Asma Ben Abacha, NLM/NIH, USA
- Laurent Besacier, Naver Labs Europe
- Yves Bestgen, F.R.S-FNRS et UCL
- Philippe Blache, LPL, CNRS
- Nathalie Camelin, LIUM, Le Mans Université
- Rémi Cardon, STL CNRS, Université de Lille
- Peggy Cellier, IRISA, INSA Rennes
- Thierry Charnois, LIPN, CNRS Université Sorbonne Paris Nord
- Vincent Claveau, IRISA, CNRS
- Maximin Coavoux, Université Grenoble Alpes, CNRS
- Mathieu Constant, ATILF, Université de Lorraine
- Benoit Crabbé, Université de Paris, LLF
- Béatrice Daille, LS2N, CNRS, Université de Nantes
- Mathieu Dehouck, CNRS, LATTICE
- Gaël Dias, Université de Normandie
- Patrick Drouin, OLST, Université de Montréal
- Emmanuelle Esperança-Rodier, LIG, Université Grenoble Alpes
- Dominique Estival, Western Sydney University
- Olivier Ferret, CEA List, Université Paris-Saclay
- Cyril Grouin, LISN, CNRS, Université Paris-Saclay
- Gaël Guibon, Télécom Paris et SNCF
- Olivier Hamon, Syllabs
- Thierry Hamon, Université Paris-Saclay, CNRS, LISN & Université Sorbonne Paris Nord
- Nabil Hathout, CLLE, CNRS

- Amir Hazem, LS2N, CNRS, Université de Nantes
- Nicolas Hernandez, LS2N, CNRS, Université de Nantes
- Stéphane Huet, LIA, Université d’Avignon
- Christine Jacquin, LS2N, CNRS, Université de Nantes
- Sylvain Kahane, Modyco, Université Paris Nanterre
- Mikaela Keller, MAGNET, Université Lille & CRISAL
- Olivier Kraïf, LIDILEM, Université Grenoble Alpes
- Matthieu Labeau, Telecom Paris
- Éric Laporte, LIGM, Université Gustave Eiffel
- Gwénolé Lecorvé, Univ Rennes, CNRS, IRISA
- Benjamin Lecouteux, LIG, Université Grenoble Alpes
- Claire Lemaire, Lairdil, Université Paul Sabatier, Toulouse III ; LIG, Université Grenoble Alpes
- Yves Lepage, Université Waseda, Japon
- Cedric Lopez, EMVISTA
- Denis Maurel, Université de Tours, Lifat
- Anne-Lyse Minard, LLL, CNRS, Université d’Orléans
- Richard Moot, LIRMM, CNRS & Université de Montpellier
- Véronique Moriceau, IRIT, Université de Toulouse
- Emmanuel Morin, LS2N, CNRS, Université de Nantes
- Luka Nerima, LATL-CUI, Université de Genève
- Aurélie Névéol, LISN, CNRS, Université Paris-Saclay
- Jian-Yun Nie, Université de Montréal
- Damien Nouvel, ERTIM, INALCO
- Sylvain Pogodalla, LORIA, INRIA
- Jean-Philippe Prost, Aix-Marseille Université et Université de Montpellier
- Solen Quiniou, LS2N, CNRS, Université de Nantes
- Christian Raymond, IRISA, INSA Rennes

- Christian Retoré, LIRMM Univ Montpellier CNRS
- Sophie Rosset, LISN, CNRS, Université Paris-Saclay
- Didier Schwab, LIG, Université Grenoble Alpes
- Pascale Sébillot, IRISA, INSA Rennes
- Gilles Sérasset, LIG, Université Grenoble Alpes
- Ludovic Tanguy, CLLE, Université de Toulouse
- Xavier Tannier, Sorbonne Université, INSERM, LIMICS
- Andon Tchechmedjiev, EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales
- Charles Teissedre, SYNAPSE
- Juan-Manuel Torres-Moreno, Laboratoire Informatique d'Avignon / UA
- Nicolas Turenne, United International College, Chine
- François Yvon, LISN, CNRS, Université Paris-Saclay
- Pierre Zweigenbaum, LISN, CNRS, Université Paris-Saclay

Co-Président.e.s RECITAL

- Amel Fraisse (GERIICO)
- Rémi Cardon (STL)

Comité de lecture RECITAL

- Jean-Yves Antoine (Université François Rabelais de Tours)
- Sonia Badene (Linagora IRIT)
- Rachel Bawden (INRIA)
- Johana Bodard (THIM/CHart EA4004 – Université Paris 8)
- Chloé Braud (IRIT – CNRS)
- Johanna Mayra Cordova (INALCO)
- Núria Gala (Aix-Marseille Université, LPL CNRS)
- Mahault Garnerin (Université Grenoble Alpes)
- Loïc Grobol (Lattice)
- William Havard (Université Grenoble Alpes)

- Laurine Huber (LORIA)
- Mikaela Keller (Université de Lille – INRIA)
- Yves Lepage (Waseda University)
- Anne-Laure Ligozat (LISN, CNRS, Université Paris-Saclay, ENSIIE)
- Damien Nouvel (INALCO)
- Patrick Paroubek (LISN, CNRS, Université Paris-Saclay)
- Thierry Poibeau (LaTTiCe-CNRS)
- Laurent Romary (INRIA & HUB-ISDL)
- Nicolas Turenne (INRA UPEM)
- Zheng Zhang (Schlumberger, AI Lab)
- Pierre Zweigenbaum (LISN, CNRS, Université Paris-Saclay)

Table des matières

I	Articles longs	1
	Auto-encodeurs variationnels : contrecarrer le problème de posterior collapse grâce à la régularisation du décodeur	2
	<i>Alban Petit, Caio Corro</i>	
	Biais de genre dans un système de traduction automatique neuronale : une étude préliminaire	11
	<i>Guillaume Wisniewski, Lichao Zhou, Nicolas Ballier, François Yvon</i>	
	Exploration des relations sémantiques sous-jacentes aux plongements contextuels de mots	26
	<i>Olivier Ferret</i>	
	La génération de textes artificiels en substitution ou en complément de données d'apprentissage	37
	<i>Vincent Claveau, Antoine Chaffin, Ewa Kijak</i>	
	Open Information Extraction : Approche Supervisée et Syntaxique pour le Français	50
	<i>Massinissa Atmani, Mathieu Lafourcade</i>	
	Plongements Interprétables pour la Détection de Biais Cachés	64
	<i>Tom Bourgeade, Philippe Muller, Tim Van de Cruys</i>	
	Transport Optimal pour le Changement Sémantique à partir de Plongements Contextualisés	81
	<i>Syrielle Montariol, Alexandre Allauzen</i>	
	Vers la production automatique de sous-titres adaptés à l'affichage	91
	<i>François Buet, François Yvon</i>	
II	Articles courts	105
	Analyse en dépendances du français avec des plongements contextualisés	106
	<i>Loïc Grobol, Benoit Crabbé</i>	
	Caractérisation des relations sémantiques entre termes multi-mots fondée sur l'analogie	115
	<i>Yizhe Wang, Béatrice Daille, Nabil Hathout</i>	
	Construire des ressources collaboratives pour les langues peu dotées : une modélisation orientée communauté	125
	<i>Elvis Mboning, Ornella Wandji</i>	
	Contribution d'informations syntaxiques aux capacités de généralisation compositionnelle des modèles seq2seq convolutifs	134
	<i>Diana Nicoleta Popa, William N. Havard, Maximin Coavoux, Eric Gaussier, Laurent Besacier</i>	
	Définition et détection des incohérences du système dans les dialogues orientés tâche.	142
	<i>Léon-Paul Schaub, Vojtech Hudecek, Daniel Stancl, Ondrej Dusek, Patrick Paroubek</i>	
	Évaluation de méthodes et d'outils pour la lemmatisation automatique du français mé-	

diéval	153
<i>Cristina Holgado, Alexei Lavrentiev, Mathieu Constant</i>	
Extraction automatique de relations sémantiques d’hyponymie et d’hyponymie dans un corpus métier	162
<i>Camille Gosset, Mokhtar Boumedyén Billami, Mathieu Lafourcade, Christophe Bortolaso, Mustapha Deras</i>	
Formalisation de la relation entre les verbes imperfectifs et perfectifs en ukrainien	171
<i>Olena Saint-Joanis, Max Silberztein</i>	
Intérêt des modèles de caractères pour la détection d’événements	179
<i>Emanuela Boros, Romaric Besançon, Olivier Ferret, Brigitte Grau</i>	
Lemmatization of Historical Old Literary Finnish Texts in Modern Orthography	189
<i>Mika Hämmäläinen, Niko Partanen, Khalid Alnajjar</i>	
Méta-apprentissage : classification de messages en catégories émotionnelles inconnues en entraînement	199
<i>Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefevre, Chloé Clavel</i>	
Prédire l’aspect linguistique en anglais au moyen de transformers	209
<i>Eleni Metheniti, Tim van de Cruys, Nabil Hathout</i>	
Sifting French Tweets to Investigate the Impact of Covid-19 in Triggering Intense Anxiety	219
<i>Mohamed Amine Romdhane, Elena Cabrio, Serena Villata</i>	
Stratégie Multitâche pour la Classification Multiclasse	227
<i>Houssam Akhmouch, Hamza Bouanani, Gaël Dias, Jose G Moreno</i>	
TREMoLo : un corpus multi-étiquettes de tweets en français pour la caractérisation des registres de langue	237
<i>Jade Mekki, Delphine Battistelli, Nicolas Béchet, Gwénoél Lecorvé</i>	
Un modèle Transformer Génératif Pré-entraîné pour le _____ français	246
<i>Antoine Simoulin, Benoit Crabbé</i>	
Une étude des avis en ligne : généralisabilité d’un modèle d’évaluation	256
<i>Hyun Jung Kang, Iris Eshkol-Taravella</i>	
III Résumés d’articles internationaux	264
Extraction d’arguments basée sur les transformateurs pour des applications dans le domaine de la santé	265
<i>Tobias Mayer, Elena Cabrio, Serena Villata</i>	
Intégration de tâches : étiquetage morpho-syntaxique, analyse syntaxique et analyse sémantique traités comme une tâche unique	268
<i>Timothée Bernard</i>	
Modéliser la perception des genres musicaux à travers différentes cultures à partir de ressources linguistiques	270

Elena V. Epure, Guillaume Salha-Galvan, Manuel Moussallam, Romain Hennequin

Revitalisation des langues autochtones via le prétraitement et la traduction automatique neuronale : le cas de l'inuktitut 273

Tan Le Ngoc, Fatiha Sadat

Simplification automatique de textes biomédicaux en français : lorsque des données précises de petite taille aident 275

Remi Cardon, Natalia Grabar

Tabouid : un jeu de langage et de culture générale généré à partir de Wikipédia 278

Timothée Bernard

Première partie
Articles longs