



AfIA

Association française
pour l'Intelligence Artificielle

TALN-RECITAL

*Conférence sur le Traitement Automatique des
Langues Naturelles*

PFIA 2019



Table des matières

Emmanuel Morin, Sophie Rosset et Pierre Zweigenbaum (TALN) Anne-Laure Ligozat et Sahar Ghannay (RECITAL).	
Éditorial	7
.	
Comités	8
 Volume I : Articles longs	
Syrielle Montariol et Alexandre Allauzen.	
Apprentissage de plongements de mots dynamiques avec régularisation de la dérive	13
Victor Connes et Nicolas Dugué.	
Apprentissage de plongements lexicaux par une approche réseaux complexes	27
Ludovic Tanguy, Pauline Brunet et Olivier Ferret.	
Comparaison qualitative et extrinsèque d'analyseurs syntaxiques du français : confrontation de modèles distributionnels sur un corpus spécialisé	39
Loïc Vial, Benjamin Lecouteux et Didier Schwab.	
Compression de vocabulaire de sens grâce aux relations sémantiques pour la désambiguïsation lexicale	55
Natalia Grabar, Cyril Grouin, Thierry Hamon et Vincent Claveau.	
Corpus annoté de cas cliniques en français	71
Antoine Caubrière, Natalia Tomashenko, Yannick Estève, Antoine Laurent et Emmanuel Morin.	
Curriculum d'apprentissage : reconnaissance d'entités nommées pour l'extraction de concepts sémantiques	85
Anissa Hamza et Delphine Bernhard.	
Détection des ellipses dans des corpus de sous-titres en anglais	99
Tim Van de Cruys.	
La génération automatique de poésie en français	113
Marco Dinarelli et Loïc Grobol.	
Modèles neuronaux hybrides pour la modélisation de séquences : le meilleur de trois mondes	127
Amalia Todirascu, Marion Cargill et Thomas Francois.	
PolylexFLE : une base de données d'expressions polylexicales pour le FLE	143
 Volume II : Articles courts	
Kate Thompson, Nicholas Asher, Philippe Muller et Jeremy Auguste.	
Analyse faiblement supervisée de conversation en actes de dialogue	159
Salima Mdhaffar, Yannick Estève, Nicolas Hernandez, Antoine Laurent et Solen Quiniou.	
Apport de l'adaptation automatique des modèles de langage pour la reconnaissance de la parole : évaluation qualitative extrinsèque dans un contexte de traitement de cours magistraux	167
Sonia Badene, Kate Thompson, Jean-Pierre Lorré et Nicholas Asher.	
Apprentissage faiblement supervisé de la structure discursive	175
Frédéric Béchet, Cindy Aloui, Delphine Charlet, Géraldine Damnati, Johannes Heinecke, Alexis Nasr et Frédéric Herlédan.	
CALOR-QUEST : un corpus d'entraînement et d'évaluation pour la compréhension automatique de textes	185
Iris Eshkol-Taravella, Mariame Maarouf, Marie Skrovec et Flora Badin.	
Chunker différents types de discours oraux : défis pour l'apprentissage automatique	195
Yuming Zhai, Gabriel Illouz et Anne Vilnat.	

Classification automatique des procédés de traduction	205
Guillaume Wisniewski.	
Combien d'exemples de tests sont-ils nécessaires à une évaluation fiable ? Quelques observations sur l'évaluation de l'analyse morpho-syntaxique du français.	215
Tsanta Randriatsitohaina et Thierry Hamon.	
De l'extraction des interactions médicament-médicament vers les interactions aliment-médicament à partir de textes biomédicaux : Adaptation de domaine	223
Fiammetta Namer, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout et Delphine Tribout.	
Demonette2 - Une base de données dérivationnelles du français à grande échelle : premiers résultats	233
Elise Bigeard et Natalia Grabar.	
Détecter la non-adhérence médicamenteuse dans les forums de discussion avec les méthodes de recherche d'information	245
Rémi Cardon et Natalia Grabar.	
Détection automatique de phrases parallèles dans un corpus biomédical comparable technique/simplifié 255	
Benoît Sagot.	
Développement d'un lexique morphologique et syntaxique de l'ancien français	265
Adrien Bardet, Fethi Bougares et Loïc Barrault.	
Étude de l'apprentissage par transfert de systèmes de traduction automatique neuronaux	275
Antoine Perquin, Gwénoélé Lecorvé, Damien Lolive et Laurent Amsaleg.	
Évaluation objective de plongements pour la synthèse de parole guidée par réseaux de neurones 285	
Sara Meftah, Nasredine Semmar, Youssef Tamaazousti, Hassane Essafi et Fatiha Sadat.	
Exploration de l'apprentissage par transfert pour l'analyse de textes des réseaux sociaux	293
Syrielle Montariol, Aina Garí Soler et Alexandre Allauzen.	
Exploring sentence informativeness	303
Fréjus A. A. Laleye, Antonia Blanié, Antoine Brouquet, Dan Benhamou et Gaël de Chalendar.	
Hybridation d'un agent conversationnel avec des plongements lexicaux pour la formation au diagnostic médical	313
Nadia Bebashina-Clairet et Mathieu Lafourcade.	
Inférence des relations sémantiques dans un réseau lexico-sémantique multilingue	323
Jean-Yves Antoine, Marion Crochetet, Céline Arbizu, Emmanuelle Lopez, Samuel Pouplin, Amélie Besnier et Mathieu Thebaud.	
Ma copie adore le vélo : analyse des besoins réels en correction orthographique sur un corpus de dictées d'enfants	333
Olga Seminck, Vincent Segonne et Pascal Amsili.	
Modèles de langue appliqués aux schémas Winograd français	343
Patricia Chiril, Farah Benamara, Véronique Moriceau, Marlène Coulomb-Gully et Abhishek Kumar.	
Multilingual and Multitarget Hate Speech Detection in Tweets	351
Iris Eshkol-Taravella et Hyun Jung Kang.	
Observation de l'expérience client dans les restaurants	361
Laurent Kevers, Florian Guéniot, A. Ghjacumina Tognotti et Stella Retali-Medori.	
Outils pour une langue peu dotée grâce au TALN : l'exemple du corse et de la BDLC	371
Amira Barhoumi, Nathalie Camelin, Chafik Aloulou, Yannick Estève et Lamia Hadrich Belguith.	
Plongements lexicaux spécifiques à la langue arabe : application à l'analyse d'opinions	381
Saoussen Mathlouthi Bouzid et Chiraz Ben Othmane Zribi.	
Q-learning pour la résolution des anaphores pronominales en langue arabe	391

Tom Bourgeade et Philippe Muller.	
Représentation sémantique distributionnelle et alignement de conversations par chat	399
Quentin Gliosca et Pascal Amsili.	
Résolution des coréférences neuronale : une approche basée sur les têtes	409
Amir Hazem, Béatrice Daille, Dominique Stutzmann, Jacob Currie et Christine Jacquin.	
Réutilisation de textes dans les manuscrits anciens	417
Aleksandra Miletić, Delphine Bernhard, Myriam Bras, Anne-Laure Ligozat et Marianne Vergez-Couret.	
Transformation d’annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l’alsacien et l’occitan	427
Yoann Dupont.	
Un corpus libre, évolutif et versionné en entités nommées du français	437
Filipo Studzinski Perotto, Fadila Taleb, Eric Trupin, Youssouf Saidali, Maryvonne Holzem, Jacques Labiche et Laurent Vercouter.	
Une approche hybride pour la segmentation automatique de documents juridiques	447

Volume III : RECITAL

Mathilde Regnault.	
Adaptation d’une métagrammaire du français contemporain au français médiéval	459
Mérimèe Bouhandi.	
Apport des termes complexes pour enrichir l’analyse distributionnelle en domaine spécialisé 473	
Jessica López Espejel.	
Automatic summarization of medical conversations, a review	487
Bruno Oberle.	
Détection automatique de chaînes de coréférence pour le français écrit : règles et ressources adaptées au repérage de phénomènes linguistiques spécifiques	499
Ygor Gallina.	
Etat de l’art des méthodes d’apprentissage profond pour l’extraction automatique de termes-clés 513	
Emmanuelle Kelodjoue.	
Extraction d’opinions pour l’analyse multicritère à partir de corpus oraux transcrits : État de l’art	525
Léon-Paul Schaub et Cyndel Vaudapiviz.	
Les systèmes de dialogue orientés-but : état de l’art et perspectives d’amélioration	541
Mathilde Veron.	
Lifelong learning et systèmes de dialogue : définition et perspectives	563
Manon Scholivet.	
Méthodes de représentation de la langue pour l’analyse syntaxique multilingue	577
Dusica Terzic.	
Parsing des textes journalistiques en serbe à l’aide du logiciel Talismane	591
Sandra Bellato.	
Vers la traduction automatique d’adverbiaux temporels du français en langue des signes française 605	

Volume IV : Démonstrations

Didier Schwab, Pauline Trial, Céline Vaschalde, Loïc Vial et Benjamin Lecouteux.	
Apporter des connaissances sémantiques à un jeu de pictogrammes destiné à des personnes en situation de handicap : un ensemble de liens entre WordNet et Arasaac, Arasaac-WN	619

Guillaume Dubuisson Duplessis, Sofiane Kerroua, Ludivine Kuznik et Anne-Laure Guénet. Cameli @ : analyses automatiques d'e-mails pour améliorer la relation client	623
Marine Schmitt, Élise Moreau, Mathieu Constant et Agata Savary. Démonstrateur en-ligne du projet ANR PARSEME-FR sur les expressions polylexicales	627
Olivier Hamon, Kévin Espasa et Sara Quispe. SylNews, un agréfilter multilingue	631
Ioan Calapodescu, Caroline Brun, Vasilina Nikoulina et Salah Aït-Mokhtar. “Sentiment Aware Map” : exploration cartographique de points d’intérêt via l’analyse de sentiments au niveau des aspects	635
Alexandre Arnold, Gérard Dupont, Catherine Kobus, François Lancelot et Pooja Narayan. Interprétation et visualisation contextuelle de NOTAMs (messages aux navigants aériens) ...	639

Éditorial

La 26^e édition de la conférence TALN et la 21^e édition de la session jeunes chercheuses et chercheurs RECITAL se déroulent cette année à Toulouse au sein de la Plateforme française d'intelligence artificielle (PFIA). TALN a une longue tradition de tenue conjointe avec des conférences de domaines proches. Cette pratique a été initiée avec les Journées d'étude sur la parole (JEP) en 2002 à Nancy puis depuis 2008 tous les quatre ans (2008 : Avignon, 2012 : Grenoble, 2016 : Paris). Elle s'est diversifiée avec la Conférence de recherche d'information et applications (CORIA) en 2018 à Rennes. Elle innove cette année avec un hébergement à Toulouse au sein de PFIA. Ces événements sont l'occasion de rencontres enrichissantes pour tous. Cette année, ce ne sont pas moins de huit conférences, sans compter les ateliers associés, aux sessions desquelles les participants à TALN-RECITAL pourront se mêler : APIA (5^e Conférence sur les Applications Pratiques de l'Intelligence Artificielle), CAp (21^e Conférence sur l'Apprentissage Automatique), IC (30^{es} Journées Francophones Ingénierie des Connaissances), JFPDA (14^{es} Journées Planification, Décision et Apprentissage), JFSMA (27^{es} Journées Francophones sur les Systèmes Multi-Agents), JIAF (13^{es} Journées d'Intelligence Artificielle Fondamentale), RJCIA (17^e Rencontre des Jeunes Chercheurs en Intelligence Artificielle), ainsi que CNIA (22^e Conférence Nationale en Intelligence Artificielle), qui regroupe les thématiques de l'intelligence artificielle non couvertes par les conférences précédentes.

Les conférences invitées plénières, les sessions de présentations affichées et de démonstrations, les déjeuners et pauses café, les dîners de la conférence sont autant de moments programmés pour que se retrouvent les participants de toutes les conférences. Nous tenons à saluer la qualité de la planification et du suivi du comité scientifique de la plateforme ainsi que le grand travail du comité d'organisation, le tout visant à assurer que l'ensemble des conférences se tiennent dans les meilleures conditions et au meilleur coût.

Pour la deuxième année consécutive, les modalités de soumission à TALN se faisaient avec un appel unique et un seul format de soumission en article court pouvant être étendu en article long sur proposition du comité de programme (et demande préalable des auteurs). Nous avons ainsi reçu soixante cinq articles courts et le comité de programme a proposé à dix articles le passage en format long (15 %) et a retenu trente et un articles en format court (48 %). Chaque article a été relu par trois membres du comité de lecture en s'appuyant le cas échéant sur des relecteurs additionnels. Le comité de programme s'est appuyé sur ces relectures pour sélectionner lors d'une réunion plénière les articles composant le programme. C'est un fonctionnement auquel nous sommes profondément attachés pour assurer une diversité dans les thématiques abordées. L'ensemble des évaluations ont été réalisées en double aveugle. Nous remercions les membres des comités de programme et de lecture (à parité femme – homme) pour leur contribution indispensable à ce processus. Le programme de la conférence est complété par quatre démonstrations sélectionnées par le comité de programme. Les titres des sessions donnent une idée des thématiques abordées par la conférence. Ils comprennent des paliers et tâches habituels du TAL (Morphologie et Syntaxe, Syntaxe, Résolution d'anaphores, Multilinguisme), reflètent la place prise par l'apprentissage (Apprentissage par transfert et modèles de langue, Plongements de mots), l'importance fondamentale que continuent à jouer les corpus et bases de données lexicales (Ressources), et l'intérêt du TAL pour des domaines particuliers (Langues spécialisées, Traitement de la langue biomédicale). Comme chaque année, l'ATALA a décerné un prix de thèse dont la récipiendaire présentera son travail en session plénière. La conférence a invité la présentation d'instruments récents du CNRS par leurs coordinatrices : d'une part le pré-GDR TAL (INS2I / informatique), qui adopte une vision inclusive du traitement de la langue (écrite, orale, signée), couvrant les communautés du traitement automatique des langues, du traitement automatique du langage parlé et de la recherche d'information ; d'autre part le GDR LIFT (INSHS / sciences du langage) sur la linguistique informatique, formelle et de terrain.

Cette année, dix-sept articles ont été soumis à RECITAL. Après avoir été chacun évalué par deux membres du comité de programme, quatre articles ont été retenus pour une présentation orale (soit un taux de sélection pour présentation orale de 24 %), et sept autres ont été retenus pour une présentation sous forme de poster (taux de sélection global de 65 %). Nous avons ainsi pu donner l'opportunité à douze jeunes chercheuses et chercheurs, en grande majorité en début de thèse, de présenter leurs travaux à la communauté. Nous remercions le comité de programme (également à parité femme – homme) pour leur minutieux travail de relecture.

Nous souhaitons pour finir au public de ces conférences une semaine riche en découvertes scientifiques et en rencontres de nouveaux collègues, dans une ambiance assurément chaude pour toute la semaine.

Emmanuel Morin, Sophie Rosset et Pierre Zweigenbaum (TALN)
Anne-Laure Ligozat et Sahar Ghannay (RECITAL)

Comités

Présidents de TALN

- Emmanuel Morin (LS2N, Université de Nantes)
- Sophie Rosset (LIMSI, CNRS, Université Paris-Saclay)
- Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay)

Membres du CP de TALN

- Delphine Bernard (LiLPa, Université de Strasbourg)
- Chloé Braud (LORIA, CNRS)
- Nathalie Camelin (LIUM, Le Mans Université)
- Peggy Cellier (IRISA, INSA Rennes)
- Benoît Crabbé (LLF, Université Paris Diderot)
- Iris Eshkol-Taravella (MoDyCo, Université Paris Nanterre)
- Cécile Fabre (CLLE-ERSS, Université Toulouse - Jean Jaurès)
- Núria Gala (LPL, Aix Marseille Université)
- Thierry Hamon (LIMSI, Université Paris Nord)
- Philippe Langlais (RALI/DIRO, Université de Montréal)
- Gwénolé Lecorvé (IRISA, Université de Rennes 1)
- Aurélie Névéol (LIMSI, CNRS, Université Paris-Saclay)
- Damien Nouvel (ERTIM, INaLCO)
- Didier Schwab (LIG, Université Grenoble Alpes)
- Xavier Tannier (LIMICS, Université Pierre et Marie Curie)

Comité de lecture de TALN

- Gilles Adda (LIMSI, CNRS, Université Paris-Saclay)
- Salah Ait-Mokhtar (Naver Labs Europe)
- Alexandre Allauzen (LIMSI, CNRS, Université Paris-Saclay)
- Maxime Amblard (LORIA, Université de Lorraine)
- Jean-Yves Antoine (LIFAT, Université de Tours)
- Loïc Barrault (LIUM, Le Mans Université)
- Denis Béchet (LS2N, Université de Nantes)
- Frederic Béchet (LIS, Aix-Marseille Université)
- Patrice Bellot (LIS, Aix-Marseille Université)
- Asma Ben Abacha (Lister Hill Center, National Library of Medicine)
- Laurent Besacier (LIG, Université Grenoble Alpes)
- Yves Bestgen (ILC, Université catholique de Louvain)
- Philippe Blache (LPL, CNRS, Aix-Marseille Université)
- Fethi Bougares (LIUM, Le Mans Université)
- Thierry Charnois (LIPN, Université Paris 13)
- Vincent Claveau (IRISA, CNRS)
- Chloé Clavel (LTCI, Télécom ParisTech)
- Kevin Bretonnel Cohen (University of Colorado School of Medicine)
- Béatrice Daille (LS2N, Université de Nantes)
- Géraldine Damnati (Orange Labs)
- Gaël Dias (GREYC, Normandie Université)
- Marco Dinarelli (LIG, CNRS)
- Patrick Drouin (OLST, Université de Montréal)
- Dominique Estival (MARCS, Western Sydney University)
- Yannick Estève (LIUM, Le Mans Université)
- Olivier Ferret (CEA LIST)
- Karën Fort (STIH, Sorbonne Université)
- Thomas Francois (CENTAL, Université catholique de Louvain)
- Éric Gaussier (LIG, Université Grenoble Alpes)
- Jérôme Goulian (LIG, Université Grenoble Alpes)

- Natalia Grabar (STL, CNRS)
- Cyril Grouin (LIMSI, CNRS, Université Paris-Saclay)
- Olivier Hamon (Syllabs)
- Nabil Hathout (CLLE-ERSS, CNRS)
- Amir Hazem (LS2N, Université de Nantes)
- Nicolas Hernandez (LS2N, Université de Nantes)
- Stéphane Huet (LIA, Université d'Avignon et des Pays de Vaucluse)
- Christine Jacquin (LS2N, Université de Nantes)
- Sylvain Kahane (Modyco, Université Paris Nanterre)
- Olivier Kraif (LIDILEM, Université Grenoble Alpes)
- Mathieu Lafourcade (LIRMM, Université de Montpellier)
- David Langlois (LORIA, Université de Lorraine)
- Eric Laporte (LIGM, Université Paris-Est Marne-la-Vallée)
- Thomas Lavergne (LIMSI, Université Paris Sud, Université Paris-Saclay)
- Joseph Le Roux (LIPN, Université Paris 13)
- Benjamin Lecouteux (LIG, Université Grenoble Alpes)
- Yves Lepage (Waseda University)
- Denis Maurel (LIFAT, Université de Tours)
- Richard Moot (LIRMM, CNRS)
- Véronique Moriceau (IRIT, Université Paul Sabatier)
- Philippe Muller (IRIT, Université Paul Sabatier)
- Alexis Nasr (LIS, Aix Marseille Université)
- Adeline Nazarenko (LIPN, Université Paris 13)
- Luka Nerima (Université de Genève)
- Jian-Yun Nie (RALI/DIRO, Université de Montréal)
- Yannick Parmentier (LORIA, Université de Lorraine)
- Sebastian Peña Saldarriaga (Dictanova)
- Thierry Poibeau (Lattice, CNRS)
- Alain Polguère (ATILF, Université de Lorraine)
- Jean-Philippe Prost (LIRMM, Université de Montpellier)
- Solen Quiniou (LS2N, Université de Nantes)
- Christian Raymond (IRISA, INSA Rennes)
- Christian Retoré (LIRMM, Université de Montpellier)
- Djamé Seddah (ALMAnaCH, Paris Sorbonne Université)
- Gilles Serasset (LIG, Université Grenoble Alpes)
- Michel Simard (NRC, Canada)
- Kamel Smali (LORIA, Université de Lorraine)
- Pascale Sébillot (IRISA, INSA Rennes)
- Ludovic Tanguy (CLLE-ERSS, Université Toulouse - Jean Jaurès)
- Juan-Manuel Torres-Moreno (LIA, Université d'Avignon et des Pays de Vaucluse)
- Guillaume Wisniewski (LIMSI, Université Paris-Sud, Université Paris-Saclay)
- François Yvon (LIMSI, CNRS, Université Paris-Saclay)

Relecteurs additionnels de TALN

- Jingshu Liu (Dictanova)
- Emile Chapuis (LTCI, Télécom ParisTech)
- Caroline Langlet (LTCI, Paris Sorbonne Université)
- Joseph Lark (Dictanova)
- Alexandre Garcia (LTCI, Télécom ParisTech)

Présidentes de RECITAL

- Anne-Laure Ligozat (LIMSI, CNRS, Université Paris-Saclay)
- Sahar Ghannay (LIMSI, CNRS, Université Paris-Saclay)

Membres du CP de RECITAL

- Jean-Yves Antoine (LIFAT, Université de Tours)

- Ismail Badache (ESPE / LIS, Aix-Marseille Université)
- Amira Barhoumi (LIUM, Université du Maine - MIRACL Sfax)
- Rachel Bawden (University of Edinburgh)
- Aurélien Bossard (LIASD, Université Paris 8)
- Chloé Braud (LORIA, CNRS)
- Nathalie Camelin (LIUM, Université du Maine)
- Rémi Cardon (STL, Lille)
- Peggy Cellier (IRISA, INSA Rennes)
- Antoine Doucet (L3i, Université de la Rochelle)
- Maha Elbayad, LIG/ Inria
- Arnaud Ferré (LIMSI-CNRS/MaIAGE-INRA, Université Paris-Saclay)
- Amel Fraisse (Gériico, Lille)
- Thomas François (CENTAL, Université catholique de Louvain)
- Nicolas Hernandez (LS2N, Université de Nantes)
- Yann Mathet (Greyc, Université de Caen)
- Alice Millour (STIH, Université Paris-Sorbonne)
- Anne-Lyse Minard (LLL, Orléans)
- Jose Moreno (IRIT, UPS)
- Tsanta Randriatsitohaina (LIMSI, Université Paris-Sud, Université Paris-Saclay)
- Loïc Vial (LIG, Université Grenoble Alpes)

Volume I : Articles longs

