



Traitement Automatique des Langues Naturelles
(TALN) ¹

Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles.
Atelier TAL et Humanités Numériques (TAL-HN)

Ludovic Moncla, Carmen Brando (Éds.)

Avignon, France, 27 juin au 1^{er} juillet 2022

1. <https://taln2022.univ-avignon.fr>

Avec le soutien de



Préface

Atelier TAL et Humanités Numériques (TAL&HN)

Ludovic Moncla¹ Carmen Brando²

(1) Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, F-69621 Villeurbanne, France

(2) EHESS, CRH, UMR8558, Paris, France

`ludovic.moncla@insa-lyon.fr`, `carmen.brand@ehess.fr`

L'atelier TAL&HN, associé à la conférence TALN | RECITAL 2022, s'inscrit dans le cadre du travail mené avec l'action de recherche Humanités Numériques Spatialisées² soutenue par le GdR CNRS MAGIS. Elle fait suite à une série d'événements organisés par notre action de recherche tels que l'atelier Humanités Numériques Spatialisées associé à la conférence SAGEO 2021 et la série d'ateliers Geospatial Humanities³ associés à la conférence internationale ACM SIGSPATIAL. Cette première édition de l'atelier associé à la conférence TALN reprend les principaux thèmes du TAL appliqués à des problématiques ou des corpus issus des travaux en sciences humaines et sociales :

- préparation de corpus et acquisition de données en SHS,
- interopérabilité de données,
- développement de ressources et Open Data,
- fouille de textes, apprentissage automatique et apprentissage profond,
- recherche d'informations, extraction de connaissances, reconnaissance d'entités nommées,
- classification de documents,
- graphe de connaissances et ontologies.

En complément, nous proposons un focus sur deux thématiques en particulier (1) l'articulation des pratiques d'encodage de corpus SHS par la norme TEI et des opportunités d'automatisation proposées par le TAL et (2) le développement de méthodes d'apprentissage (automatique ou profond) et l'entraînement de modèles de langue pour le traitement de corpus SHS (historique, littéraire, etc.).

Notre comité de programme est constitué d'une équipe pluridisciplinaire (TAL et humanités numériques) de 26 collègues chercheurs et enseignants-chercheurs issus de diverses équipes et laboratoires de recherche francophones. Nous avons reçu 13 soumissions présentant des résultats de recherche (à des niveaux différents de maturité) en lien avec les thèmes de l'atelier. Chaque article a été relu par au moins 3 relecteurs du comité de programme et suite à la phase de relecture 10 ont été acceptés pour publication. Ces 10 articles se regroupent en 3 sessions autour des thèmes suivants : 1) l'étude du genre dans des corpus de textes, 2) l'annotation sémantique en tenant compte des erreurs d'OCéRisation et 3) l'analyse de discours et sémantique des textes. Du point de vue de la science ouverte et des normes dites FAIR, les corpus utilisés dans ces travaux sont publiés en TEI (Text Encoding Initiative) et parfois déposés dans des entrepôts de données de la recherche comme Ortolang d'Huma-Num.

Les articles présentés proposent un aperçu des travaux en humanités numériques qui s'appuient sur l'utilisation et le développement de méthodes numériques pour le traitement et l'analyse de corpus textuels. Nous tenons à remercier les auteurs pour leurs contributions ainsi que les participants à l'atelier et nous espérons que ce numéro sera utile pour la communauté.

2. <https://aphns-magis.projet.liris.cnrs.fr/>

3. <https://ludovicmoncla.github.io/sigspatial-geohumanities-2022/>

Table des matières

Exploration orientée entités : étude du genre dans le Mercure de France	1
<i>Yoann Dupont, Marguerite Bordry</i>	
Flux d'informations dans les systèmes encodeur-décodeur. Application à l'explication des biais de genre dans les systèmes de traduction automatique.	10
<i>Lichao Zhu, Guillaume Wisniewski, Nicolas Ballier, François Yvon</i>	
LDAPol : vers une méthodologie de contextualisation des discours politiques	19
<i>Jeanne Vermeirsche, Eric Sanjuan, Tania Jiménez</i>	
Les animaux chinois de Buffon : identification automatique des jugements critiques dans l'Histoire naturelle (1749-1789)	28
<i>Axel Le Roy, Motasem Alrahabi, Glenn Roe</i>	
Reconnaissance automatique des appellations d'œuvres visuelles antiques	36
<i>Aurore Lessieux, Iris Eshkol-Taravella, Anne-Violaine Szabados, Marlène Nazarian</i>	
Reconnaissance d'entités nommées sur des sorties OCR bruitées : des pistes pour la désambiguïsation morphologique automatique	45
<i>Caroline Koudoro-Parfait, Gaël Lejeune, Richy Buth</i>	
Réinterroger l'édition numérique et la consultation d'œuvres anciennes : traçabilité, accessibilité, interprétabilité	56
<i>Emmanuel Giguët, Julia Roger</i>	
Romanciers et romancières du XIXème siècle : une étude automatique du genre sur le corpus GIRLS	66
<i>Marco Naguib, Marine Delaborde, Blandine Andrault, Anaïs Bekolo, Olga Seminck</i>	
Simulation d'erreurs d'OCR dans les systèmes de TAL pour le traitement de données anachroniques	78
<i>Baptiste Blouin, Benoit Favre, Jeremy Auguste</i>	
TAL et Littérature comparée. Détection automatique des correspondances textuelles entre les réécritures d'un mythe	88
<i>Karolina Suchecka, Nathalie Gasiglia</i>	