Laboratoire d'informatique de Paris Nord

LIPN

Rapport d'activité 2017-2022 et projet 2023-2028

Équipe RCLN : Représentation des Connaissances et Langage Naturel

cnrs

UNIVERSITÉ SORBONNE PARIS NORD

Couverture : © Olivier Bodini, « Twitter 7 mars », mars 2023.

Partie I — **Le LIPN** — **page 5**

The LIPN at a glance — 7

- Scientific activities
- Environment
- Doctoral studies

Partie II — **Equipe RCLN** — **page 13**

Informations générales sur l'équipe RCLN — 15

- Membres de l'équipe RCLN
- Les thématiques scientifiques et leurs enjeux
- Profil d'activités liées à la recherche
- Prise en compte des recommandations du précédent rapport

Introduction du portfolio — 27

Auto-évaluation de l'équipe RCLN — 29

- Domaine 2. Attractivité
- Domaine 3. Production scientifique
- Domaine 4. Inscription des activités de recherche dans la société

Trajectoire de l'équipe RCLN — 53

- SWOT
- Trajectoire

Partie III — **Acronymes** — **page 61**

Partie I
Le LIPN

Chapitre 1
The LIPN at a glance

The LIPN (Laboratoire d'Informatique de Paris-Nord) was created in 1985, has been associated with CNRS since January 1992, before becoming a CNRS UMR in January 2001. The LIPN is the computer science laboratory of University Sorbonne Paris Nord (USPN) which is part of the Alliance Sorbonne-Paris-Cité ASPC.

The laboratory comprises 5 teams with more than 70 permanent staff (full-time researchers, professors and associate professors or research engineers), 9 technical and administrative staff, as well as PhD students and postdocs. It has overall more than 150 members most of the time. After having experienced impressive growth during the last two five-year contracts, the number of teacher-researchers at LIPN has decreased, weakening some of its areas of expertise.

The scientific activities display many collaborations, be they national, international, or through research projects, in particular with the industry. LIPN members are extremely involved on the national scene, in national boards (at CNRS, CNU — national universities council, HCERES, etc.), and in institutional cooperation projects (ANR). On the international scene, members are widely present in editorial boards, programme committees, expert pools of different countries, international schools, and conduct collaborative research with colleagues worldwide.

The LIPN at a glance

1.1 Scientific activities

The laboratory is structured into 5 teams and a transverse axis, focussing on the following main areas.

- **A³ (Apprentissage Artificiel et Applications): Machine Learning and Applications**

The A³ team tackles machine learning problems and covers a wide spectrum of issues, ranging from supervised and unsupervised learning to reinforcement learning. Its research is fed, coordinated and evaluated thanks to various applications in the field of pattern recognition and data mining. Research in A³ focuses on the following main topics: algebraic and logical models of learning, collaborative and transfer learning, cluster analysis, dimensionality reduction, link prediction in social networks, recommender systems, and learning structures from complex data.

- **AOC (Algorithmes et Optimisation Combinatoire): Algorithms and Combinatorial Optimisation**

The AOC team develops research in optimisation on graphs, mathematical programming, parallel and distributed computing. Optimisation on graphs is conducted with a particular focus on complexity, polyedral theory and approximation. The team has expertise in design and analysis of mathematical programming approaches and algorithms, and develops both exact and heuristic approaches for solving linear and nonlinear problems. Many distributed environments issues are also considered: distributed middleware and architectures, distributed programming and distributed algorithms.

- **CALIN (Combinatoire, ALgorithmique et INteractions): Combinatorics, ALgorithmics, and INteractions**

The CALIN team brings together researchers with skills in a variety of aspects of combinatorics (analytic, bijective, geometric, and algebraic). They are interested in the complexity of algorithms and in the fine-grained analysis of data structures. They also study problems of physics with a distinct combinatorial flavour, or apply methods from physics to combinatorial problems. The team is organised into two intersecting subgroups that reflect these interests: one of them focuses on the analysis of algorithms and combinatorial structures, and the other is devoted to the interactions between combinatorics and physics. Geometry, both as a tool and a subject of study is transversal to these two subgroups.

- **LoVe (Logique et Vérification : Logic and Verification**

The LoVe team tackles different aspects of computational and software models, from theory to applications. It is organised into two subgroups. The logic, programming and complexity group studies the relationship between logic and programming languages, especially for what concerns denotational semantics of programs and proofs, under the Curry-Howard correspondence. It has a special expertise in the quantitative aspects of programming, including the links with computational complexity. It also hosts experts in formal proofs. The verification group develops model checking approaches for concurrent systems, in particular in a parametric, distributed and timed setting. The research spans from fundamental aspects to application domains, including an open verification platform software development.

- **RCLN (Représentation des Connaissances et Langage Naturel): Knowledge Representation and Natural Language**

The RCLN team carries out innovative work for the analysis and exploration of text corpora as well as for the acquisition of knowledge from texts or knowledge graphs. The team is particularly active in problems that are little studied in language, such as the treatment of under-endowed or specialised languages, and is working on approaches to deal with the complexity of problems in texts or knowledge graphs, such as syntactic analysis, information extraction, or graph completion using methods from deep learning, combinatorial optimisation, data mining or inductive logic programming.

- **The transverse axis Complexités: Complexities**

The *Complexities* research axis spreads across the whole laboratory and brings together all those people whose work is related, directly or indirectly, with computational complexity. It is primarily composed of members of the LoVe, CALIN and AOC teams, but its themes potentially intersect also with the other two teams of LIPN. The activities of the axis are centered around a weekly seminar.

1.2 Environment

The LIPN has many research cooperations with its environment: other laboratories at University Sorbonne Paris Nord, but also in the Alliance Sorbonne-Paris-Cité (ASPC), with industry, *etc.*

- **MathSTIC research federation, and University Sorbonne Paris laboratories**

Research collaborations with other laboratories of University Sorbonne Paris Nord have been favoured over the years, in particular with LAGA (UMR 7539, mathematics laboratory), L2TI (UR3043, signal and image processing laboratory).

This led to the creation of a new research federation in Mathematics and Information Technology MathSTIC, which became a CNRS Research Federation – FR 3734 – in January 2016), gathering members of LAGA, LIPN, and L2TI laboratories, in order to enhance cross-fertilisation in four of their major areas of expertise:

— Optimisation and learning applied to digital contents;
— High-performance computing, distributed systems;
— Mathematical physics, statistical physics, combinatorics;
— Categories: between calculation and topology.

- **Regional setting: DIM RFSI and EFL LabEx**

All LIPN teams strongly collaborate with other groups in the Île-de-France region. Furthermore, LIPN has been heading the regional initiative in computer science, DIM RFSI, which fostered regional collaborative projects between different laboratories and companies in computer science or application domains. It also allowed for setting up a large regional cluster infrastructure for HPC and deep learning.

The team RCLN at LIPN is a major actor in the EFL LabEx (Empirical Foundations of Linguistics). Its multidisciplinarity nature is a key to evolution of research topics at the meeting point of linguistics and computer science.

- **International cooperations**

Members of the laboratory have many collaborations worldwide, with exchanges of researchers. Moreover, approximately 15 foreign researchers are invited for a period of two to fours weeks at our laboratory each year to carry collaborative research. Cooperation projects involve laboratories in Denmark, Netherlands, Norway, Poland, Italy,

Tunisia, Marocco, Canada, United States, Argentina, Brazil, Chile, Mexico, Uruguay, Singapore, Taïwan, Vietnam, Australia.

- **Industrial collaborations**

Most research areas of the LIPN also participate in collaborations with the industry, mainly through research projects (ANR, FUI, industrial collaborations) and CIFRE PhDs. The LIPN is involved in a joint laboratory (LabCom) on satellite image processing. Furthermore some researchers of the LIPN have also created a start-up company, or benefit from a few months industrial experience. Moreover, the LIPN is involved in different boards of 2 business clusters (pôles de compétitivité) which favour technology transfer and research projects with companies at a regional level.

1.3 Doctoral studies

The doctoral studies are organised within the doctoral school ED146 *Galileo : Science, Technology, Health*, which addresses three major scientific areas, LIPN being part of the first group:

1. mathematics, computer science, signal processing;
2. physics, materials, engineering sciences;
3. health sciences, medicine, human biology, chemistry, ethology.

The doctoral school organises mandatory courses, and allocates scholarships to doctoral students. Every year, approximately five of them are allocated to the LIPN (out of 25 for the whole doctoral school).

- **PhDs and Habilitations**

Over the 2017–2022 period, 62 PhD defences took place. Half of the doctoral students were financed by a scholarship from the doctoral school, and the others through research projects, industry, or international support.

During the same period, 11 habilitations (Habilitations à diriger les Recherches) were delivered.

- **Math & Computer Science Graduate School**

The M&CS Graduate school is an international 5-year high-level curriculum founded for 8 years. It is endorsed by the Galileo Doctoral School and three local masters: Mathematics, Computer Science, Engineering and Innovation in Images and Networks. The LIPN and the two other laboratories of MathSTIC research federation actively participate in this graduate school. The interactions between Mathematics and Computer Science are highlighted in the graduate school to promote interdisciplinarity in teaching and research.

- **Masters courses**

A masters programme in computer science is delivered at University Sorbonne Paris Nord. It currently features two specialities:

— EID^2 (Exploration Informatique des Données et Décisionnel): Machine Learning and Data Science and its international curriculum Data Science;

— PLS (Programmation et Logiciels Sûrs): Programming and Secure Software.

Both specialities lead to either industrial or research careers, depending on the master thesis.

LIPN researchers also deliver courses in other masters programmes of universities in the parisian region.

Furthermore, they are involved in international cooperations for setting up masters programmes in foreign countries, e.g. Madagascar, Vietnam.

- **Summer schools and conferences**

LIPN supports the organisation and participation in thematic schools, by encouraging doctoral students attendance (which is recognised as mandatory courses), talks by members of the laboratory, as well as financial support to the organisation of such events.

In particular, Column Generation School 2018, Mathématiques expérimentales: méthodes et pratiques (MathExp) 2018, Caleidoscope research school 2019, Deep Learning & Data science thematic school every two years, were co-organised by LIPN members. Some of these thematic schools take place in University Sorbonne Paris Nord, while others are elsewhere in France or abroad.

Partie II
Equipe RCLN

Overview

For almost three decades, research carried out by the RCLN team has focused on Natural Language Processing (NLP) and knowledge representation. Knowledge is generally transmitted through natural language, and thus Knowledge Representation, a major issue in Artificial Intelligence, Cognitive Science and Semantic Web, is concerned with processing the knowledge expressed in natural language. In a complementary way, text understanding may involve knowledge and reasoning.

The RCLN team brings together skills in automatic language processing, corpus linguistics, semantic web, as well as in data mining and machine learning. These complementary skills give it an original positioning and allow it to carry out innovative work for the analysis and exploration of text corpora as well as for the acquisition of knowledge from texts or knowledge graphs. The team is particularly active in problems that are little studied in language, or complex problems such as the treatment of under-endowed or specialised languages, the detection of neologisms, the analysis of microblogs or the generation and evaluation of abstractive summaries. The team is also working on approaches to deal with the complexity of problems in texts or knowledge graphs, such as syntactic analysis, simultaneous extraction of semantic relations and named entities, or graph completion using methods from deep learning, combinatorial optimisation, data mining or inductive logic programming. The team is structured around three closely related research areas: Algorithms for syntactic and semantic analysis of texts, Exploration and generation of texts and Knowledge Acquisition. These three axes are complementary: the syntactic-semantic analysis of corpora can serve as a basis for text exploration and knowledge acquisition, and conversely the analysis algorithms can take advantage of the knowledge acquired.

These researches have led to results published in the major NLP conferences (EMNLP, NAACL, COLING) and in the fields of semantic web and knowledge acquisition (Semantic Web, KBS, ISWC, ESWC, K-CAP), as well as Data Mining (IDA). They have also led to an important software development: about thirteen software tools are publicly available.

The RCLN research team is strongly involved in the LabEx scientific project "Empirical Foundations of Linguistics" (EFL), in which we play different key coordination roles.

For the next five years, the team will focus on three poles of innovative research: syntactic and semantic analysis taking advantage of deep learning methods and combinatorial optimisation, ressources and analysis methods for under-endowed languages, new usages of large language models (LLM) that are based on knowledge graphs, and knowledge graphs enrichment methods. Taking advantage of the results obtained so far, the team will continue its research on its federative project: scientific literature mining.

Chapitre 1

Informations générales sur l'équipe RCLN

Sommaire

1.1 Membres de l'équipe RCLN 16
1.1.1 Membres permanents 16
1.1.2 Membres non permanents au 01/01/2023 17
1.1.3 Doctorants associés au 01/01/2023 17
1.1.4 Évolution de l'équipe (permanents) 18
1.2 Les thématiques scientifiques et leurs enjeux 18
1.2.1 Algorithmes pour l'analyse syntaxique et sémantique des textes . . 19
1.2.2 Exploration et génération de textes 20
1.2.3 Acquisition de connaissances à partir de textes et de graphes de connaissances 21
1.2.4 Axe transverse concernant la fouille de littérature scientifique . . . 23
1.3 Profil d'activités liées à la recherche 24
1.4 Prise en compte des recommandations du précédent rapport . . . 24

Nom de l'équipe : RCLN Représentation des Connaissances et Langage Naturel (RCLN)

Responsable de l'équipe : Nathalie Pernelle

1.1 Membres de l'équipe RCLN

Informations générales sur l'équipe RCLN

Corps	2017	2018	2019	2020	2021	2022	(2023)
DR	0	0	0	0	0	0	(0)
CR	0	0	0	0	0	0	(0)
PU	3	2	2	3	3	3	(3)
MCF	10	10	10	10	10	10	(10)
Prag	0	0	0	0	0	0	(0)
IR	2	2	2	2	2	2	(2)
Post-docs et ATER	2	1	1	1	1	1	(0)
Ingénieurs (CDD)	2	0	0	0	1	1	(0)
Doctorants	11	10	5	5	7	8	(7)
Total	*30*	*25*	*20*	*21*	*24*	*25*	*(22)*

1.1.1 Membres permanents

(membres ou, *en italique, anciens membres*, et, **en gras, responsable de l'équipe** au 01/01/2023)

	Nom	*Prénom*	*Situation*	*Institution*
	Audibert	Laurent	MCF	IUTV
	Buscaldi	Davide	MCF	IUTV
	Cartier	Emmanuel	MCF	UFR LLSHS
	Charnois	Thierry	PU	IUTV
	Garcia Flores	Jorge	IR	CNRS
	Gayral	*Françoise*	MCF	IUTV
	Gangemi	Aldo	PU	IG
	Grezka	*Aude*	*IR*	*CNRS*
	Le Roux	Joseph	MCF	IG
	Lévy	François	PU émérite	IUTV
	Nazarenko	Adeline	PU	IG
	Pernelle	**Nathalie**	PU	IG
	Recanati	Catherine	MCF	IG

	Rozenknop	Antoine	MCF	IG
	Salotti	Sylvie	MCF	IUTV
	Tomeh	Nadi	MCF	IG
	Zargayouna	Haïfa	MCF	IUTV
	Zarrouk	Manel	MCF	IG

1.1.2 Membres non permanents au 01/01/2023

	Nom	*Prénom*	*Situation*	*Institution*
	Beugnet	Vincent	CIFRE	Thales DMS
	El Khbir	Niama	Contrat Doctoral	LabEx
	Felhi	Ghazi	Allocataire	USPN
	Floquet	Nicolas	Allocataire	USPN
	Liyanage	Vijini	Allocataire	USPN
	Schulz	Alexandre	Contrat Doctoral	Huawei
	Zaratiana	Urchade	CIFRE	Initiativas

1.1.3 Doctorants associés au 01/01/2023

(*Les doctorants associés sont encadrés par des membres du LIPN et dépendent d'une école doctorale extérieure à l'Université Paris 13.*)

Nom	*Prénom*	*Situation*	*Institution*
Abdine	Hadi	Contrat doctoral	Ecole Polytechnique
Aleman Manzanarez	Gerardo	Contrat doctoral	Tec de Monterrey
Annanda	Sousa	Contrat doctoral (SFI)	Univ. Galway
Azevedo	Lucas	Contrat doctoral (SFI)	Univ. Galway
Boukhatem	Nabil Moncef	CIFRE	Ecole Polytechnique
Khajeh Nassiri	Armita	Contrat doctoral	Paris Saclay
Kulic	Ammar	Allocation Doctorale	USPN (ED Erasme)
Lovera	Fernando	Salarié	Univ. Simon Bolivar
Simonne	Lucas	Contrat doctoral	Paris Saclay
Wozniak	Weronica	Cotutelle	Université de Lodz

1.1.4 Évolution de l'équipe (permanents)

Les effectifs de l'équipe RCLN sont restés stables si l'on considère le début et la fin de la période. Le recrutement d'une professeure des Universités en décembre 2019 et d'une maître de conférence en janvier 2020 a permis de compenser un départ en retraite et un départ en disponibilité de longue durée. Ce renouvellement a permis de renforcer les compétences de l'équipe en Acquisition de connaissances et Graphes de connaissances. L'arrivée d'une IR en linguistique en septembre 2018 a permis d'accroître nos compétences en linguistique de corpus. Cependant, il faut noter l'engagement d'Adeline Nazarenko (PU) au CNRS en tant que DAS sur toute la période et la mise en disponibilité d'un an d'un MCF pour raisons personnelles. Ainsi, l'équipe a été particulièrement affectée par l'ensemble de ces départs et engagements en 2018 et 2019.

Informations générales sur l'équipe RCLN

Départs

- **Sept 2017** A. Gangemi, PU IG (disponibilité)
- **Sept 2018** N. Tomeh, MCF IG (disponibilité)
- **Mars 2019** F. Gayral, MCF IUTV (retraite)

Arrivées

- **Sept 2018** A. Grezka, IR, CNRS (mutation du LDI)
- **Sept 2019** N. Tomeh, MCF IG (retour de disponibilité)
- **Dec 2019** N. Pernelle, PU IG (MCF Paris Saclay)
- **Jan 2020** M. Zarrouk, MCF IG (Senior Postdoc Univ. Galway)

1.2 Les thématiques scientifiques et leurs enjeux

L'équipe RCLN (Représentation des Connaissances et Langage Naturel) réunit des compétences en traitement automatique des langues, linguistique de corpus, web sémantique, ainsi qu'en fouille de données et en apprentissage automatique. Ces compétences complémentaires lui confère un positionnement original et permet de mener des travaux novateurs pour l'analyse et l'exploration des corpus de textes ainsi que pour l'acquisition de connaissances à partir de textes ou de graphes de connaissances. L'équipe se positionne en particulier sur des problèmes peu étudiés en langue ou difficiles comme le traitement de langues sous-dotées, les langues de spécialité, la détection de néologismes sémantiques, l'analyse de microblogs ou encore la génération et l'évaluation de résumés abstractifs. L'équipe s'attache également à définir des approches permettant de gérer la complexité des problèmes dans les textes ou les graphes de connaissance, tels que l'analyse syntaxique , l'extraction simultanée de relations sémantiques et d'entités nommées, ou encore la découverte de règles expressives dans les graphes en faisant appel à des méthodes issues de l'apprentissage profond, de l'optimisation combinatoire, de la fouille de données ou encore de la programmation logique inductive.

L'équipe est fortement impliquée dans l'animation et les travaux du LabEx EFL *Empirical Foundations of Linguistics*, et en particulier dans l'axe 5 *Analyse sémantique computationnelle*. Adeline Nazarenko a été co-directrice du labex depuis sa création jusqu'en 2022, co-direction aujourd'hui assurée par Thierry Charnois qui assure également la coordination de l'axe 5 et Emmanuel Cartier a co-piloté les activités de formation jusque février 2023. L'équipe participent activement à la préparation d'un nouveau Labex qui sera soumis début 2024.

L'équipe est structurée autour de trois axes de recherche fortement liés : *Algorithmes pour l'analyse syntaxique et sémantique des textes*, *Exploration et génération de textes* et enfin *Acquisition de connaissances à partir de de textes ou de graphes de connaissances*. Ces trois axes sont complémentaires : l'analyse syntaxico-sémantique de corpus peut servir de base à l'exploration de texte et à l'acquisition de connaissances,

et réciproquement les algorithmes d'analyse peuvent tirer parti des connaissances acquises.

1.2.1 Algorithmes pour l'analyse syntaxique et sémantique des textes

Dans cet axe, l'équipe développe des approches d'analyse syntaxique et sémantique en s'appuyant sur des techniques issues de l'optimisation combinatoire et de l'apprentissage automatique, en particulier de l'apprentissage profond.

- **Analyse Syntaxique**

L'objectif de l'analyse syntaxique de dépendances basée sur les graphes est de découvrir la structure la plus adéquate pour une phrase d'entrée en calculant un score pour toutes les analyses candidates possibles et en renvoyant celle qui a obtenu le score le plus élevé. Cependant, le nombre d'analyses candidates est exponentiel. L'équipe a développé des techniques d'analyse syntaxique pour le TAL fondées sur l'apprentissage profond et les méthodes issues de l'optimisation dans les graphes telles que la relaxation lagrangienne, décomposition, algorithme de Frank-Wolfe, ou encore la montée par bloc de coordonnées pour obtenir des résultats à l'état de l'art en analyse syntaxique [**CI-27**, **CI-11**, **CI-5**]. Ces méthodes, développées pour certaines d'entre elles en collaboration avec l'équipe AOC, permettent de combiner des contraintes permettant de garantir la bonne formation des structures produites, avec des réseaux de neurones permettant de paramétrer les analyseurs à partir de corpus d'exemples. Les représentations en dépendances peuvent également encoder certaines informations sémantiques de la phrase qui peuvent varier selon les jeux de données. Nous avons tiré partie de deux jeux de données disponibles pour l'arabe dont l'un représente des informations syntaxiques et morpho-syntaxiques et l'autre relativement plus focalisé sur les relations sémantiques. Nous avons proposé un algorithme multitâche d'analyse syntaxique des dépendances, basé sur des LSTM hiérarchiques de type easy-first, capable de décoder une phrase dans plusieurs formalismes. Nous avons montré qu'un tel système conjoint est plus performant que le système de référence mono-tâche sur les banques de dépendances du CATiB et de l'UD Arabic et que les informations linguistiques disponibles dans chaque formalisme aident à faire de meilleures prédictions pour l'autre [**CI-15**]. Enfin, l'équipe a également défini des méthodes de génération non supervisées, en particulier pour le désenchevêtrement (*i.e. disentanglement*) des rôles syntaxiques et de la sémantique lexicale, particularité essentielle pour les travaux cherchant à obtenir des méthodes interprétables. Ces méthodes sont fondées sur les auto-encodeurs variationnels et le mécanisme d'attention croisée [**CI-34**].

- **Analyse sémantique**

En ce qui concerne l'analyse sémantique des textes, nous nous intéressons à la fois à la tâche d'extraction d'entités nommées, et à l'extraction de relations ou d'évènements.

Pour la tâche d'extraction d'entités nommées, nous avons notamment proposé une méthode de filtrage des espaces de recherche de modèles CRF semi-Markovien qui permet leur passage à l'échelle et l'amélioration de leur performance [CI-28]. La classification de relations consiste à trouver les instances des relations sémantiques, c'est-à-dire les ensembles d'entités, et à classer la relation qui les lie selon une typologie existante. Dans un cadre non supervisé, les types de relations doivent être déduits directement des données. Dans ce cadre, nous avons montré comment des plongements de mots issus de modèles pré-entrainés pouvaient être utilisés pour identifier des types de relations sémantiques en utilisant une mesure de similarité relationnelle originale basée sur la combinaison des vecteurs d'entrée et de sortie CBOW de word2vec [**CI-4**].

L'équipe s'est également focalisée sur la thématique d'extraction d'information spécifique à la langue arabe. Nous avons ainsi développé une collaboration étroite avec le Camel Lab (Univ. de NY, Abu Dabi) pour concevoir des systèmes adaptés à la complexité de cette langue. Les travaux existants se concentrent principalement sur la reconnaissance d'entités nommées et peu de travaux ont été réalisés sur l'extraction de relations et la reconnaissance d'événements en arabe, ce qui constitue une tâche complexe en raison de la richesse morphologique et des idiosyncrasies de cette langue. L'équipe a développé des modèles basés sur de l'apprentissage profond supervisé permettant d'extraire, à partir d'un texte, le graphe d'entités avec leurs relations et les événements avec leurs arguments [CO-60]. Pour atteindre les résultats de l'état de l'art sur ces tâches, nous nous appuyons sur l'analyse morphologique pour répondre aux difficultés spécifiques à l'arabe et utilisons un algorithme de recherche en faisceau différentiable pour permettre l'apprentissage *end-to-end* des paramètres du modèle.

Informations générales sur l'équipe RCLN

Enfin, nous avons proposé la première tâche sur l'extraction de relations sémantiques à la compétition internationale SemEval 2018. Le défi se concentre sur les relations sémantiques spécifiques au domaine scientifique et comprend trois sous-tâches différentes qui permettent de comparer et quantifier l'effet de différentes étapes de prétraitement sur les résultats de la classification des relations. Cette tâche a attiré un total de 32 participants avec 158 soumissions aux différents scénarios. Les jeux de données ont depuis été utilisés par de nombreuses approches : le track est cité par 173 publications [**CO-23**].

1.2.2 Exploration et génération de textes

Cet axe de recherche réunit des travaux qui cherchent à combiner analyse des textes et connaissances pour classifier, annoter, résumer ou générer des documents.

- **Exploration de textes**

L'exploration des microblogs fait l'objet de nombreuses recherches car ceux-ci peuvent être utilisés pour identifier des tendances, détecter de nouveaux évènements, ou pour prédire des phénomènes physiques et sociaux. L'équipe a abordé le problème de la géolocalisation de microblogs en combinant des architectures neuronales avec des bases de connaissances pour extraire des informations contextuelles, même en l'absence de références géographiques directes [CI-13]. L'algorithme, appelé Sherloc, exploite les toponymes contenus dans le message, extrait leur sémantique d'une ressource géographique et les intègre dans un espace métrique qui capture la distance sémantique entre eux. Les toponymes les plus proches sémantiquement d'un message sont identifiés et regroupés en fonction de leur emplacement spatial. Sherloc ne nécessite aucun entrainement préalable et peut déduire la localisation dans une ville avec une grande précision. D'autres travaux sur les microblogs ont porté sur l'analyse de sentiments [CO-51], ou encore la détection de trolls [CO-45].

- **Génération de textes**

Nous nous sommes intéressés à doter les robots de capacités en langage naturel leur permettant de raconter leur expérience dans l'exécution d'une certaine tâche. Ces capacités narratives peuvent être utilisées pour surveiller l'activité d'un robot dans un environnement autonome ou améliorer la qualité de la communication avec des interlocuteurs. Contrairement aux travaux antérieurs sur la verbalisation des robots, qui se concentrent exclusivement sur sa navigation spatiale, ce travail considère les mouvements et actions réalisées, les lieux où la tâche a été exécutée et ses résultats. Nous avons défini un modèle de dialogue narratif, un graphe de situations, qui peut être

instancié en utilisant l'analyse de l'historique des tâches du robot et ses éléments significatifs et qui est utilisé pour générer une phrase et ses connecteurs linguistiques [RI-8].

L'équipe s'est également intéressée à la génération automatique de résumés abstractifs et à leur évaluation. Nous avons proposé un modèle génératif auto-régressif basé sur une architecture neuronale Transformer de type BART spécifique à la langue arabe [**CO-58**] . Ce modèle a atteint l'état de l'art sur plusieurs jeux de données de caractéristiques différentes. L'équipe a également contribué aux méthodes d'évaluation de résumés automatiques. ROUGE est sans doute la métrique la plus largement utilisée, mais celle-ci se base sur les chevauchements lexicaux entre les résumés candidats et les résumés de référence et n'est donc pas adaptée aux résumés abstractifs. Aussi, nous avons développé une méthode permettant d'évaluer la qualité des résumés abstractifs appelée GeSERA, qui généralise la méthode SERA définie pour les données médicales, et qui s'affranchit de la présence de chevauchements lexicaux en utilisant des vecteurs de similarité sémantiques basés sur Wikipédia [CI-25].

Les modèles de génération de texte excellent aujourd'hui dans la production de textes qui paraissent être générés par un humain. Ces outils peuvent être utilisés à mauvais escient et générer par exemple des textes comportant de fausses informations. Proposer des outils capables de distinguer un texte généré automatiquement d'un texte écrit par un humain devient essentiel. Nous avons proposé un jeu de données constitué d'articles originaux et d'articles générés automatiquement mais partiellement modifiés. Ces derniers permettent de simuler une utilisation plus réaliste des outils de génération automatique par rapport aux jeux de données existants dans lesquels les textes sont générés complètement automatiquement [CI-31]. Les résultats préliminaires montrent que les systèmes existants ont moins de facilité pour repérer des textes de ce type.

1.2.3 Acquisition de connaissances à partir de textes et de graphes de connaissances

Dans cet axe, nous nous intéressons aux approches permettant d'exploiter des corpus de textes ou des Graphes de connaissances volumineux pour construire ou enrichir des ressources langagières structurées ou des graphes de connaissances.

- **Construction de ressources langagières**

La ressource lexicale publique Morfetik dont nous sommes responsables permet d'obtenir, pour tout mot français, toutes ses formes (féminin et pluriel, formes conjuguées des verbes, etc.), ou, à l'inverse, d'identifier le mot (i.e. "lemme") correspondant à toute forme fléchie [RI-15]. Cette ressource, construite de manière semi-automatique en collaboration avec des linguistes, décrit aujourd'hui plus de 1000 000 formes grâce à l'introduction d'expressions et de mots composés en 2022. L'interface offre également la possibilité de consulter la ressource interne Néoveille [RE-3], notre plateforme de repérage et de suivi des néologismes en corpus dynamique et d'obtenir ainsi pour le mot recherché une visualisation des contextes dans lequel le mot apparait ainsi que des indicateurs métalinguistiques (pays, journal, domaine). L'équipe a en effet poursuivi ses travaux sur l'identification et la caractérisation des innovations lexicales dans un cadre multilingue, en développant des outils de gestion de corpus dynamiques et en diachronie, de fouille et de suivis lexicaux mobilisant différentes approches, qu'il s'agisse de méthodes d'identification supervisées à base de ressources lexicales, de méthodes de détection des événements rares sur la base de patrons syntaxiques prototypiques ou encore de modèles de langue contextuels [CL-9, RI-9]. L'intérêt s'est porté essentiellement sur les innovations lexicales dites formelles, par identification des procédés de dérivation et de composition.

- **Construction et enrichissement de Graphes de Connaissances**

En collaboration avec quatre universités européennes (Univ. Cagliari, Univ. Seville, Open Univ. of Milton Keynes, Univ. of Milano Bicocca), l'équipe RCLN a défini l'approche SCICERO [RI-20] qui permet de générer des graphes de connaissances à grande échelle, sémantiquement riches et précis à partir de la littérature scientifique, tels que CS-KG pour l'informatique (41 millions d'énoncés extraits de 6,7 millions de résumés) [CI-35], ou AI-KG [**CI-16**]. Cette approche peut être combinée avec l'approche de complétion de graphes SciCheck qui incorpore un ensemble de caractéristiques et d'heuristiques conçues pour capturer les connaissances scientifiques [RE-6] et qui est plus précise que l'état de l'art des approches basées sur les plongements ou sur des méthodes classiques de TAL. Le grand nombre d' articles scientifiques publiés chaque année, implique une difficulté croissante à suivre les nouveaux thèmes des différents domaines scientifiques et leur importance. Nous avons également proposé une stratégie de construction automatique de taxonomies de thèmes permettant d'exploiter des millions de résumés d'articles scientifiques en nous appuyant sur des modèles simples qui sont plus faciles à évaluer automatiquement, mais en améliorant nos modèles grâce à l'application d'un grand nombre d'itérations [CI-36].

Les données scientifiques constituent également l'une des sources de connaissances pertinente pour la recherche d'experts et de leur expertise. Les méthodes existantes de recherche d'experts dans les publications ne tiennent pas compte de la validation implicite par les pairs exprimés par les liens de collaborations. En collaboration avec l'équipe A3, nous avons proposé une méthode originale qui extrait l'expertise et les liens de collaborations d'un corpus de publications scientifiques annoté sémantiquement par le biais d'une ontologie, représente les informations extraites sous la forme de graphes bipartites attribués, et enfin énumère des motifs fermés apparaissant dans des noyaux (i.e Core closed pattern mining) en introduisant des critères de sélection originaux adaptés à la recherche d'experts et à leur expertise [CI-21, CO-20].

L'équipe RCLN s'est également intéressée aux approches de fouille de règles dans les graphes de connaissances. Ces règles peuvent venir enrichir le graphe de connaissance et être utilisées pour rendre explicite les régularités dans les données, compléter ou intégrer les données, détecter des données erronées, ou concevoir des systèmes hybrides. Dans les graphes de connaissances, les approches de découverte de règles doivent faire face au volume de données et à l'hypothèse du monde ouvert (OWA) qui ne permet pas de disposer facilement de contrexemples. Nous avons étudié les différences théoriques et expérimentales entre différentes sémantiques de règles de liage [RI-12], qui varient en fonction de la façon de prendre en compte l'hypothèse du monde ouvert. Dans le cadre du projet PSPC AIDA, nous avons collaboré avec le LISN pour concevoir le système RE-Miner qui découvre efficacement des règles de liage basées sur des expressions référentielles expressives représenter sous forme de motifs de graphes [**CI-14**]. Ces motifs caractérisent un individu de manière unique et nous avons montré sur différents jeux de données réels que cette approche surpassait les approches de liage à base de règles existantes (approche arrivée 1ère en terme de F-Mesure pour un track OAEI 2020) [CO-42]. Nous avons également montré que les liens erronés générés par ces approches automatiques pouvaient être en partie détectés grâce à une approche uniquement basée sur les communautés pouvant être détectées dans le graphe formé par les liens d'identité [RI-13] (collaboration avec la Vrije Univ. d'Amsterdam). De même, nous avons collaboré avec l'équipe A3 pour définir une approche nommée CRA-Miner qui permet de découvrir des règles de prédiction de la présence d'amiante dans les éléments de construction pour le CSTB (Centre Scientifique et Technique du Bâtiment) [CI-24], en s'inspirant d'approches top-down issues de la Programmation Logique Inductive mais en offrant la possibilité de prendre en compte des informations temporelles et les liens de subsomption de manière efficace. Enfin, peu d'approches s'intéressent à la découverte de règles causales dans les graphes de connaissances. Nous avons collaboré avec l'INRAE de Montpellier pour définir les différentes variantes du système DICARE

(Dicare-E, Dicare-C et Dicare E) qui découvrent des règles différentielles causales représentées en logique du premier ordre, en se basant sur le cadre du potenital outcome framework où la valeur du contrefactuel est estimée [**CI-20**, CI-29]. Le système a été utilisé pour expliquer pourquoi certaines personnes veulent réduire leur consommation de viande plus que d'autres en utilisant un graphe de connaissance représentant des données collectées à l'INRAE .

1.2.4 Axe transverse concernant la fouille de littérature scientifique

Les publications de recherche ou les ensembles de données partagées devraient être mieux exploités par des systèmes intelligents pour soutenir et accélérer les efforts scientifiques, en facilitant la recherche d'experts, la génération d'état de l'art, la génération de nouvelles hypothèses scientifiques, ou pour soutenir ou contredire certaines affirmations. Nous avons défini la fouille de littérature scientifique comme un axe transverse de l'équipe RCLN depuis 2017. Ces textes emploient une (des) langue(s) de spécialité et introduisent différentes difficultés : (1) le nombre de publications scientifiques croit de manière exponentielle et cela rend les possibilités d'analyse manuelle de plus en plus difficile ou parcellaire (e.g. le nombre d'études publiées sur le changement climatique a été par exemple multiplié par 100 depuis 1990), (2) le vocabulaire y est particulièrement évolutif que ce soit pour exprimer des entités désignant des méthodes, des outils, des organisations ou des objets d'étude, (3) les relations et entités sont souvent ambigües, difficiles à segmenter et enfin la recherche d'expert pour un domaine de spécialité se complique également avec le volume des données et la qualité variable des publications. Comme cela apparaît dans les descriptions des axes présentées ci-dessus, nous avons défini différentes approches qui portent sur la fouille de littérature scientifique et qui peuvent être employées de manière complémentaires. Ces approches vont de l'extraction de relations à la création de taxonomies ou de graphes de connaissances à partir de résumés d'articles scientifiques, en passant par une énumération et une caractérisation des experts de domaine prenant en compte une forme de validation par les pairs. De plus, nous avons constitué des ressources permettant à d'autres approches de tester leur capacité à traiter des textes de littérature scientifique (i.e. track SemEval) et nous avons construit et mis à disposition des graphes de connaissances générés à partir de millions de résumés de publications du domaine de l'informatique (CS-KG) et plus spécifiquement pour le domaine de l'intelligence artificielle (AI-KG).

1.3 Profil d'activités liées à la recherche

Activités (répartir 100 points sur ces 7 items)	
Administration et animation de la recherche : pilotage de la recherche (VP, direction d'institut, DAS, par exemple), participation à des instances d'évaluation (CNU, CoNRS, CSS, Hcéres, par exemple), responsabilité de dispositifs Idex ou Isite, direction de projets (ANR, Horizon Europe, ERC, CPER, PIA, France 2030, par exemple), responsabilités éditoriales dans des revues ou collections nationales et internationales.	40
Aide aux politiques publiques et expertise technique : pouvoirs publics aux niveaux européen, national et régional, entreprises, instances internationales comme FAO, OMS, etc.	0
Contribution à l'adossement d'enseignements innovants à la recherche : EUR, SFRI, etc.	0
Dissémination de la recherche : partage de connaissances avec le grand public, médiation scientifique, interface sciences et société	5
Recherche et encadrement de la recherche.	45
Valorisation, transfert, innovation.	20
Autres activités. (à préciser en une ligne maximum).	0

Informations générales sur l'équipe RCLN

1.4 Prise en compte des recommandations du précédent rapport

Dans cette section, nous présentons les actions entreprises pour tenir compte des recommandations HCERES données à la précédente évaluation.

— Manque de collaborations intra-équipe : pour favoriser les collaborations intra-équipe, nous avons défini un projet transverse concernant la fouille de littérature scientifique, collaboré pour constituer ou participer à des tracks tels que la campagne d'évaluation SemEval 2017, SemEval 2018, le Défi national DEFT 2019, qui implique à chaque fois deux ou trois membres permanents, incité les maîtres de conférences à participer au co-encadrement des thèses dirigées par des habilités et enfin 5 membres collaborent pour définir et implémenter la plateforme ChêneTal qui doit permettre aux linguistes d'accéder et de combiner certains des outils développés au sein de l'équipe. Ainsi, sur cette période, plus d'un tiers des articles de conférences internationales ont parmi leurs auteurs 2 ou 3 membres permanents RCLN.

— Manque de collaborations internationales : sur cette période 6 thèses ont été co-encadrées avec des laboratoires étrangers. Sur les 21 revues internationales publiées, 17 ont des co-signataires membres de laboratoires étrangers, de même sur 36 conférences internationales, 13 sont effectuées avec des collaborateurs étrangers.

— Manque de projets internationaux : nous avons pu augmenter le nombre de projets internationaux portés par l'équipe en incitant les enseignants chercheurs de l'équipe à inviter des chercheurs étrangers (8 enseignants chercheurs invités sur la période), en formaliser ces collaborations internationales via des projets nationaux tels que des projets émergence permettant de financer des missions, puis via des projets bi-latéraux, pour ensuite participer au montage de projets plus conséquents. Sur la période précédente, seulement 2 projets bilatéraux avaient été menés (Ecos-Nord et AUF-PIRAT). Sur cette période, un projet émergence implique le Camel Lab (Trauma), et 5 projets bilatéraux ont vu le jour (4 projets PHC et 1 projet Ecos-Nord). Nous sommes également partenaires d'un projet Horizon CL-23 qui vient d'être déposé en 2023.

— Risque concernant la diminution des ressources suite à une fin possible du Labex EFL : la fin du Labex EFL aurait eu en effet pour conséquence une importante diminution des ressources de l'équipe RCLN. Le Labex a été reconduit jusque 2024 et l'équipe s'implique fortement cette fois encore pour que le Labex soit renouvelé en 2024. En parallèle, nous avons augmenter nos soumissions à des appels à projet et nos collaborations industrielles ce qui a conduit à une augmentation des thèses Cifres et des projets nationaux auxquels nous participons.
— Manque d'HDR : nous sommes conscients que le nombre d'HDR est trop faible compte tenu de la taille de l'équipe : un maître de conférence a soutenu son habilitation en 2018 en linguistique, mais seulement les trois membres PU sont habilités en informatique au 1er janvier 2023 dont une qui est extrêmement prise par ses responsabilités au sein du CNRS. Nous avons incité les enseignants chercheurs à co-encadrer les thèses menées au sein de l'équipe ainsi que des thèses co-encadrées avec d'autres laboratoire et deux maîtres de conférences sont actuellement en train d'écrire leur manuscrit en vue de soutenir leur habilitation avant la fin de l'année.
— Durée des thèses trop élevée : nous avons été plus attentifs à la durée des thèses qui dépassait les 4 ans en moyenne sur la période précédente. Nous avons veillé à une bonne intégration des doctorants au sein de l'équipe, et en particulier à leur participation active au groupe de lecture et aux séminaires de l'équipe. La durée a pu être ramenée à 43,9 mois sur cette période.
— Vie de l'équipe non explicitée : L'équipe se réunit mensuellement, permanents et non-permanents, pour discuter des appels à projet, des évènements liés au Labex, des besoins en poste de permanents ou d'ATER, des recrutements de docteurs et post-doctorants, et des besoins en équipement. Un groupe de lecture bi-mensuel a été mis en place depuis la dernière période, qui est animé par Joseph Leroux et Nadi Tomeh, et les séminaires sont organisés par Davide Buscaldi. L'équipe est représentée au conseil de laboratoire par deux membres de l'équipe, et par trois membres au comité d'expert pour la section 27.

Chapitre 2
Introduction du portfolio

L'équipe RCLN propose 5 portfolios : (1) Higher-Order Dependency Parsing for Arc-Polynomial Score Functions via Gradient-Based Methods and Genetic Algorithm, (2) Traitement automatique de la langue arabe (3) Generating Knowledge Graphs from Scientific Literature (4) Campagne d'évaluation en extraction et classification de relations sémantiques dans les papiers scientifiques, et enfin (5) Morfetik, ressource morphologique.

Les portfolios 1 et 3 ont été choisis afin de présenter deux contributions scientifiques de l'équipe en analyse syntaxique et en acquisition de connaissances à partir de texte, contributions qui sont au coeur des travaux de l'équipe dans deux axes différents, et qui ont été valorisées dans des conférences et journaux de haut niveau.

Le portfolio concernant le traitement automatique de la langue arabe est un exemple permettant d'illustrer nos travaux actuels sur les langues sous-dotées, recherches que nous menons en collaboration avec le Camel Lab.

Les portfolios 4 et 5 sont des exemples de ressources que nous partageons avec la communauté scientifique. Il s'agit tout d'abord de la première campagne internationale d'évaluation sur l'extraction de relations dans les textes scientifiques, campagne qui a impliqué de nombreux participants et dont les ressources sont réutilisées jusqu'à aujourd'hui par les approches d'extraction pour évaluer leurs résultats. Enfin, le portfolio 5 présente une ressource morphologique qui est surtout à destination des linguistes. Morfetik permet d'accéder à l'ensemble existant le plus complet des formes simples et

composées du français, avec leurs variations.

Chapitre 3

Auto-évaluation de l'équipe RCLN

Sommaire

3.1 Domaine 2. Attractivité 30

3.1.1 Référence 1. L'équipe est attractive par son rayonnement scientifique et s'insère dans l'espace européen de la recherche. 30

3.1.2 Référence 2. L'équipe est attractive par la qualité de sa politique d'accompagnement des personnels. 32

3.1.3 Référence 3. L'équipe est attractive par la reconnaissance de ses succès à des appels à projets compétitifs. 33

3.1.4 Référence 4. L'équipe est attractive par la qualité de ses équipements et de ses compétences techniques. 34

3.2 Domaine 3. Production scientifique 35

3.2.1 Référence 1. La production scientifique de l'équipe satisfait à des critères de qualité. 35

3.2.2 Référence 2. La production scientifique de l'équipe est proportionnée à son potentiel de recherche et correctement répartie entre ses personnels. 35

3.2.3 Publications 36

3.3 Domaine 4. Inscription des activités de recherche dans la société . 51

3.3.1 Référence 1. L'équipe se distingue par la qualité et la quantité de ses interactions avec le monde non-académique. 51
3.3.2 Référence 2. L'équipe développe des produits à destination du monde culturel, économique et social. 52

3.1 Domaine 2. Attractivité

3.1.1 Référence 1. L'équipe est attractive par son rayonnement scientifique et s'insère dans l'espace européen de la recherche.

Le rayonnement scientifique est visible à travers les invitations dans les colloques internationaux ou nationaux. Les membres de l'équipe sont aussi très régulièrement sollicités pour participer à des comités de programme des conférences majeures au niveau international ou pour l'organisation de manifestations scientifiques dans les domaines du TAL, du web sémantique, ou encore de la linguistique, ce qui témoigne de leur reconnaissance dans ces domaines. Il se traduit aussi par une forte implication dans le pilotage de la recherche notamment au niveau national (CNRS) ou local (LabEx EFL).

Ces éléments sont détaillés ci-dessous.

Invitations dans des colloques ou conférences Les membres de l'équipe sont régulièrement invités à présenter leurs travaux dans des colloques, conférences ou ateliers de niveau international ou national :

Auto-évaluation de l'équipe RCLN

— Conférence CNRIA, Thies (Sénégal), 15 au 17 juin 2022 ;
— Spanish Conference on Information Retrieval *CERI 2018* à Saragosse (Espagne) du 25 au 27 juin 2018 ;
— 7th International Symposium on Language & Knowledge Engineering *LKE 2019*, Dublin (Irlande), octobre 2019 ;
— Workshop international *DeepOntoNLP 2020*, Bolzano (Italie) le 16 septembre 2020 ;
— Webinaire *DOING network*, LIFO, Université d'Orléans, le 17 juin 2021 ;
— Journée d'étude à l'Université de Niigata (Japon), 6 mars 2017 ;
— Journée d'étude à l'Université de Tokyo (Japon), 8 mars 2017 ;
— Journée d'étude à l'Université de Nagoya (Japon), 16 décembre 2017 ;
— Séminaire de l'Université de technologie de Chiba, Chiba-kodai (Japon), 26 octobre 2018 ;
— Colloque International La Perception : langue, discours, cognition, 6 et 7 décembre 2019, Sorbonne Université ;
— Webinaire de l'École Doctorale des Pays de Vysegràd, Université Jagellonne de Cracovie et Institut de Néophilologie de l'Université Pédagogique de Cracovie, le 23 et 24 septembre (2021).

Responsabilités éditoriales

— Comité éditorial et de rédaction du périodique Buscila-infos (Bulletin d'information en Sciences du langage), jusqu'en 2019 ;
— Comité scientifique de la collection Dixit Grammatica, L'Harmattan (Paris). Revue internationale électronique de grammaire, linguistique et langue françaises, depuis 2015 ;
— Comité de rédaction de la revue Traitement automatique du langage (3 membres de l'équipe sur la période).

Participation à des comités de programme de conférences internationales importantes. L'équipe est très investie dans les PC (comités de programme) des conférences internationales de premier plan, ainsi que dans les principales conférences nationales du domaine.

- 4 Area Chair (COLING 2020, ACL 2020, COLING 2022, CoNLL 2022)
- Action Editor pour ACL Rolling Review (depuis 2021)
- plus de 40 PC ou comités scientifiques de conférences internationales (en tant que membres) : IJCAI 2019, IJCAI-ECAI 2018,EMNLP-IJCNLP 2019,NAACL 2019,ACL 2020,AACL-IJCNLP 2020,EMNLP 2020,ACL 2021,NAACL 2021,EACL 2021,COLING 2022,CONLL 2022, EMNLP 2022, ECML-PKDD 2016 et 2017, LREC 2016, 2018, 2020, 2022, ESWC 2017, 2019, 2020, ISWC 2017, 2022, WebConf 2017, 2018, 2019, 2020, 2021, 2022, BioNLP-OST 2019, WSDL 2022, EKAW 2020,21,22, ESWC 2020,21,22, ICCS 2020, 21, 22 ;
- PC de conférences nationales : TALN, EGC, IC, CORIA, CNRIA

Organisation de conférences ou ateliers internationaux

- 3 Co-chairs de la tâche 7 (Semantic Relation Extraction and Classification in Scientific Papers) du challenge Semeval 2018 ;
- 2 Co-chair du track Knowledge Management & Ontology engineering de la conférence IEEE AICCSA (2017, 2018) ;
- 2 comités de pilotage SemEval 2020 (en tant que membre) ;
- 4 co-organisations du workshop "Deep Learning for Knowledge Graphs" (DL4KG2020) associé à ESWC ou ISWC (2019, 2020, 2021, 2022) ;
- 4 co-organisations du workshop "Sentic Computing, Sentiment Analysis, Opinion mining and Emotion Detection", satellite workshop at ESWC, (2018) ;
- 1 co-organisation du workshop "Interactions between Data Mining and Natural Language Processing" associé à ECML-PKDD, (2017) ;
- 5 co-organisations (Publication Chair) du workshop "Arabic Natural Language Processing", (2017, 2019, 2020, 2021, 2022) ;
- 1 co-organisation du workshop "Processing Language Variation : Digital Armenian (DigitAm)" associé à LREC (2022) ;
- 2 co-organisation du workshop "Language Technology for Equality, Diversity, Inclusion" associé à EACL 2021 et ACL 2022", (2021, 2022) ;
- 1 co-organisation du congrès international FRASEOLOGÍA Y TRADUCCIÓN EN HISPANOAMÉRICA, Université Ricardo Palma, (2018).

Responsabilités dans les sociétés savantes

- 1 membre du comité de pilotage (Advisory Committee) de l'ACL Special Interest Group on Arabic Natural Language Processing (depuis 2022) ;
- 1 vice-présidence de l'ATALA - Association pour le Traitement Automatique des Langues - (2017 à 2021) ;
- 2 membres du collège SIC de l'AFIA (depuis 2015 et depuis 2018) ;
- 1 membre du comité d'orientation de l'ARIA (depuis 2016) ;
- 1 membre du comité d'administration de l'ATALA (2017- 2021) ;
- 1 membre du comité de pilotage de la conférence nationale IC (depuis 2018).
- 1 membre du comité de pilotage collège HLT AFIA (depuis 2019).

Participation à des instances de pilotage ou évaluation de la recherche L'équipe est fortement investie dans les instances de pilotage de la recherche avec des responsabilités scientifiques ou administratives importantes notamment au niveau du CNRS et du LabEx EFL, ainsi que dans de multiples instances d'évaluation telles que l'HCERES, l'ANR.

Participation à des instances de pilotage

— CNRS : Direction adjointe scientifique INS2I CNRS (2016-2022) ;
— Responsabilité de la mission pour l'expertise scientifique du CNRS (2022) – puis Direction de l'institut INS2I depuis février 2023 – ;
— Co-direction du Laboratoire d'excellence "Fondements Empiriques de la Linguistique", LabEx EFL (depuis 2011) ;
— Responsabilité de l'axe 5 "Analyse sémantique computationnelle" du LabEx EFL (depuis 2011) ;
— Responsabilité de la formation du LabEx EFL (2018-2022) ;
— 2 Membre élus de la Commission Recherche et du Conseil Académique de l'université Sorbonne Paris Nord (2016 – 2020) et (2021-2022) ;
— Membre élu du Conseil d'Administration de l'université Sorbonne Paris Nord (2018-2020) ;
— Co-responsable axe "Accès à l'information et fouille de textes" du GDR TAL (depuis 2018).

Participation à des instances d'évaluation de la recherche

— Membre extérieur Conseil Scientifique LabEx « Les passés dans le présent » (depuis 2019) ;
— Membre extérieur Conseil Scientifique du LIRMM (depuis 2022) ;
— Comité HCERES (1 présidence et 1 membre expert), évaluation Laboratoire Informatique d'Avignon, (2022)
— Membre du comité d'évaluation scientifique CE23 Intelligence Artificielle de l'ANR, Appel à projets générique (2018, 2020 et 2022) ;
— Experte pour les campagnes d'avancement des IT CNRS, commission régionale d ?interclassement de la BAP D (circonscription Ile de France Villejuif) ;
— Membre du comité d'évaluation des chercheurs du CIRAD Montpellier commission ACACONGUA (2022).

Auto-évaluation de l'équipe RCLN

En outre, les membres ont été impliqués dans les responsabilités liées à l'enseignement :

— Vice présidente du département d'informatique de l'Institut Galilée (depuis 2021) ;
— Cheffe du département GEA de l'IUT de Villetaneuse (2016–2021) ;
— Responsable du BUT GEA (depuis 2021).

3.1.2 Référence 2. L'équipe est attractive par la qualité de sa politique d'accompagnement des personnels.

Sur la période considérée, l'équipe a vu l'arrivée d'une ingénieure de recherche CNRS en sciences du langage (début 2018) par mutation interne, d'une maîtresse de conférences (décembre 2019) et d'une professeure (début 2020), toutes deux en section 27 du CNU. En revanche, l'équipe peine à attirer des candidats sur les postes CR CNRS, difficultés qui peuvent s'expliquer par un vivier réduit et surtout du fait des perspectives offertes dans l'industrie, sur les domaines du traitement des langues et de l'apprentissage, c'est-à-dire de l'intelligence artificielle, un domaine, actuellement et depuis quelques années, particulièrement actif et d'actualité.

Postdoctorants, doctorants et ingénieurs. Sur la période on compte 19 doctorant(e)s dont 7 en cours, 11 soutenances (au 31 décembre 2022) et un abandon. L'équipe a su remarquablement diversifier l'origine des financements de ces thèses. Les allocations ministérielles représentent 42% (8 thèses), le reste des financements se composent de 2 allocations LabEx EFL, 1 double culture (ex COMUE SPC), 1 ANR, 1 FUI, 3 financements extérieurs (Huawei, CEA et IRSTEA) et 3 CIFRE. La durée moyenne des thèses sur la période est de 43,9 mois dont 2 thèses qui ont été menées difficilement à terme et dont la durée a dépassé 4 ans. La situation sanitaire dûe à la pandémie en 2020 et 2021

a clairement eu un impact sur cette durée. En effet, des doctorants se sont retrouvés confinés dans des petits studios durant plusieurs mois, et les contacts par visio que nous avons maintenus n'ont pas remplacé une présence en laboratoire. Cependant comparée à la dernière période, la durée moyenne des thèses s'est nettement améliorée : elle était de plus de 4 ans dans la période précédente. À l'issue de leur soutenance un doctorant a effectué un postdoc à l'étranger et a été recruté comme Maître de conférence (il a été aussi classé à l'issue du concours CR CNRS). Tous les autres autres doctorant(e)s ont été recrutés dans l'industrie. On peut aussi noter que la quasi-totalité des doctorants (18 sur 19) sont co-encadrés par un MCF, ce qui permet aux non-HDR de participer à cette activité.

Les membres de l'équipe participent aussi aux co-encadrements de doctorant(e)s d'une école doctorale hors Galilée dans le cadre de co-directions ou de co-encadrements de thèses dont 6 à l'étranger (Vénézuela, Mexique, 3 Irlande et Pologne), celle de Pologne étant en co-tutelle, auxquels il faut ajouter 2 thèse à LIX, 3 à Orsay et 1 à l'école doctorale Erasme à l'USPN. Sur la période sont ainsi recensés 12 doctorant(e)s hors école doctorale Galilée dont 10 en cours au 31 décembre 2023.

L'équipe a par ailleurs recruté 4 postdoctant(e)s dont 3 femmes, sur des financements LabEx, FUI, et projet IDEX Neoveille. Deux postocs ont ensuite obtenu un poste de Maître de Conférences.

L'équipe est particulièrement attentive au suivi et au devenir des personnels contractuels recrutés (doctorant(e), postocs et ingénieurs sur CDD). Elle les accompagne dans leurs démarches, les conseille, relis leur dossier et les prépare à des auditions dans le cas d'un concours. Ces personnels sont aussi intégrés dans l'équipe sous forme d'invitations aux réunions d'équipe, d'invitations à présenter leurs travaux en séminaire ou encore à participer aux groupes de lecture organisés régulièrement (à un rythme bimensuel ou mensuel). Les doctorant(e) et postdocs sont aussi financés par l'équipe sur projets ou par le laboratoire pour présenter leurs travaux publiés dans les conférences, et sont encouragés à participer à une école d'été. L'équipe bénéficie de deux ingénieurs de recherche permanents qui font partie de la DevTeam au niveau du laboratoire et apporte leur support technique à tous les personnels de l'équipe. Les deux IR participent aux travaux scientifique de l'équipe. Arrivée au début de la période (janvier 2018) par mutation interne, une IR a récemment été promue en classe exceptionnelle, ce qui est un fait notable et témoigne de la qualité de son intégration.

Auto-évaluation de l'équipe RCLN

Invitations de chercheurs. L'équipe invite régulièrement des chercheurs provenant de laboratoires internationaux pour développer des collaborations scientifiques ou dans le cadre de co-encadrements de doctorant(e). Ces invitations sont financées sur projets, sur fonds propres du LIPN ou par l'Université Sorbonne Paris Nord. La liste ci-dessous donne les invitations d'au moins 2 semaines et est limitée aux invités internationaux :

- Yudith Cardinale, Univeristé Simon Bolivar, Vénézuela, 2019 ;
- Raoul Monroy, TEC de Monterrey, Mexique, 2022 ;
- Adam Wyner, Swansea University, UK, 2017 et 2022 ;
- Franseco Osborne, Open University, Milton Keynes, UK, 2018 et 2019 ;
- Mustafa Jarrar, Bir Zeit University, Palestine, 2018 ;
- Nizar Habach,New York University Abu Dhabi, UAE, CamelLab, Abou Dhabi, 2018 ;
- Fernando Perez-Tellez, Tecnicol University Dublin, Ireland, 2022 ;
- Emiliano Sanchis, Universitat Politècnica de València, Espagne, 2021.

3.1.3 Référence 3. L'équipe est attractive par la reconnaissance de ses succès à des appels à projets compétitifs.

Les résultats de l'équipe en matière de réponse à des appels à projets sont concluants :

- Au niveau international, 4 projets bilatéraux PHC (2 Polonium, 1 Barrande, 1 Galilée) 25k€ au total, 1 ECOS-Nord ;

- Au niveau PIA :
  - 1 projet IDEX ex COMUE Sorbonne Paris Cité, Neoveille (2015-2017), 211k€ ;
  - 1 LabEx EFL (2011-2014), plus de 470 k€ depuis 2017 pour l'équipe ;
- Au niveau national :
  - 1 ANR PARSITI (2016–2021), 134k€ ;
  - 1 ANR DALIH (2021–2025), 130k€ ;
  - 2 DGLFLF (Délégation Générale à la Langue française et aux Langues de France) Ministère de la Culture, (2017–2018 et 2017–2020), 64k€ au total ;
  - 1 émergence TRAUMA CNRS (2020), 10k€.

  On peut aussi considérer le projet FUI PCUv2 avec des industriels et les financements CIFRE (voir en section 3.3.1).

Sur la période, l'ensemble de ces projets a ainsi permis de financer 5 thèses, 4 postdoctorant(e)s et 2 ingénieurs d'étude, et plus larement de dynamiser les activités de l'équipe, comme le montre le financement LabEx.

En effet, l'équipe RCLN est fortement impliquée dans le pilotage et les travaux du LabEx EFL dont l'ensemble des activités de recherche s'inscrivent dans le projet scientifique du LabEx. Dans ce cadre, plusieurs opérations financées par le LabEx et portées par des membres de RCLN ont clairement contribué à renforcer les thématiques de l'équipe par le biais de recrutements de postdoctorants (trois sur la période financés à 100%) et doctorants (trois dont deux financés entièrement), de missions, d'équipement (serveur de calcul GPU), de bourses de mobilité entrante (5 invitations de doctorants à l'international) et de soutiens de colloques (2 ateliers internationaux, DL4KG 2019 2020) - l'ensemble de ces financements s'élevant à plus de 470 k€ pour la période considérée (366k€ engagés et 104k€ obtenus et à engager)-. Ce contexte a également favorisé le développement de collaborations au sein du LabEx mais aussi à l'international. La visibilité du LabEx a aussi permis à l'équipe d'attirer de très bons candidats pour les recrutements doctoraux et postdoctoraux.

Auto-évaluation de l'équipe RCLN

3.1.4 Référence 4. L'équipe est attractive par la qualité de ses équipements et de ses compétences techniques.

À l'initiative de l'équipe RCLN, le laboratoire a développé depuis 2014 une infrastructure de calcul dédiée à l'intelligence artificielle et au traitement automatique des langues. Aujourd'hui, cette infrastructure met à disposition des chercheur·euse·s 5 serveurs de calcul avec des accélérateurs GPU, pour un total de 140 coeurs, 1,2 To de RAM et 26 GPUs (8 A40, 11 Quadro RTX 5000, 3 RTX2080 et 4 RTX2080Ti). Cette infrastructure rend service à une cinquantaine d'utilisateur·rices appartenant au LIPN, le L2TI, le LabEx EFL ainsi qu'à des collaborateur·rice·s internationaux·ales. L'infrastructure permet de mettre au point les paramètres pour entraîner des grands modèles de langues (LLM) avant d'éventuellement continuer l'entraînement sur des infrastructures nationales, comme Jean-Zay. Les serveurs sont administrés par un IR de l'équipe RCLN et une IE de l'USPN. Quant aux démos et systèmes mis en ligne par l'équipe, le serveur (80 coeurs et 640 Go de RAM) sert de vitrine à la mise en ligne des logiciels Neoveille (détection de néologismes), Morfetik (ressource morphologique) et SDMC (extraction de motifs). Il hébergera aussi la plateforme de mise en ligne de chaînes de traitements ChêneTAL, en cours de développement. Ce serveur est ouvert aussi aux collaborateur·rice·s internationaux·ales de l'équipe RCLN, et il compte plus de 120 utilisateur·rices inscrit·e·s.

3.2 Domaine 3. Production scientifique

	2017	2018	2019	2020	2021	2022	2023	*Total*
Livres	0	1	0	0	0	0	0	*1*
Chapitres de livre	3	3	1	3	1	0	0	*11*
Revues internationales	7	2	2	5	2	3	0	*21*
Autres revues	0	1	2	0	2	2	0	*7*
Conférences internationales	7	3	3	4	9	10	0	*36*
Autres communications	18	14	6	10	4	10	2	*64*
Total	*35*	*24*	*14*	*22*	*18*	*25*	*2*	*140*

3.2.1 Référence 1. La production scientifique de l'équipe satisfait à des critères de qualité.

Comme le montrent ses publications, l'équipe est reconnue pour ces travaux en analyse syntaxique et sémantique qui s'appuient sur des méthodologies basées sur de l'apprentissage profond mais aussi sur de l'optimisation combinatoire. Elle est également reconnue pour son expertise concernant la construction ou l'enrichissement de ressources langagières et de graphes de connaissances à partir de corpus ou de données structurées. Dans ces domaines, les conférences internationales sont le moyen de publication privilégié par la communauté particulièrement en traitement automatique des langues, ce qui explique le nombre relativement plus important de publications dans ces supports par rapport aux revues internationales.

Les membres de l'équipe ont ainsi publié dans les conférences les plus réputées de ces domaines : EMNLP, NAACL, AACL, Coling, EACL, CoNLL et ISWC pour les publications réputées de rang A. Les publications dans les revues sont aussi de très bon niveau notamment les revues classées Q1 dans scimagojr : Semantic Web, Knowledge Based Systems, Journal of Geographical Information Science, Future Generation Computer Systems, et celles classées Q1 ou Q2 selon les années : Open Linguistics, Data intelligence, Big Data Research.

Auto-évaluation de l'équipe RCLN

Signalons l'article SemEval 2018 qui porte sur l'organisation par l'équipe en collaboration avec des chercheurs allemands du premier challenge sur l'extraction et la classification de relations sémantiques dans le domaine de la littérature scientifique. Cet article n'est pas classé dans les conférences internationales mais a une visibilité importante notamment en raison des jeux de données textuelles que nous avons mis à disposition : il a fait l'objet à ce jour de 173 citations (source Google Scholar). Ce travail est décrit dans le portfolio.

3.2.2 Référence 2. La production scientifique de l'équipe est proportionnée à son potentiel de recherche et correctement répartie entre ses personnels.

Le nombre d'équivalents temps plein sur la période pour les membres permanents et tenant compte de la présence partielle de certains membres est $35,75$. On obtient :

— $21/35,75 = 0,58$ article de revue internationale par personne et par an ;
— $32/35,75 = 0.90$ article de conférence internationale par personne et par an.

Si un article est compté le nombre de fois égal au nombre de membres de l'équipe ou du laboratoire l'ayant co-signé, on obtient les nombres suivants :

— $0,62$ article de revue internationale par personne et par an, soit sensiblement le même nombre que dans le cas précédent (1 seul article avec 2 signatures de membres de l'équipe) ;

— $1,31$ article de conférence internationale par personne et par an. 11 articles de conférences ont 2 membres parmi les signataires, et 2 articles en ont 3.

À la différences des articles de conférence pour lesquels il est donc fréquent que plusieurs membres publient ensemble, les revues sont plus fréquemment co-signées par un seul membre de l'équipe mais avec des collaborateurs étrangers. S'agissant des conférences internationales sélectives, il est aussi intéressant de noter la bonne dynamique de publication qui a été enclenchée durant la période : le nombre de publications dans cette catégorie est passé de 3 en 2018 et 2019, à 9 en 2021, puis à 10 en 2022.

Les doctorants participent également à la production scientifique de l'équipe et ont tous publié durant la préparation de leur thèse au moins 2 articles dont au moins un au niveau international (6 doctorants ont soutenu en 2017 ou 2018 et ont des publications antérieures qui n'apparaissent pas dans ce document). 1 article de revue internationale 11 articles de conférence internationale ont été signés par les doctorants, le plus souvent avec leur(s) encadrant(s), dont certains dans de très bonnes conférences (NAACL, AACL, IWPT, EMNLP) sans compter les publications des 6 doctorants ayant publié antérieurement à 2017). Les postdoctorant(e)s accueillis plus de 6 mois ont également publié durant leur séjour dans l'équipe (EMNLP, CiCling).

Un petit nombre de membres permanents (environ 16% sur la période) ne participent que peu ou pas à la production scientifique de l'équipe. Il s'agit pour l'essentiel d'un investissement dans des tâches lourdes d'enseignement et de responsabilités liées à l'enseignement à l'IUT. S'ils le souhaitent, l'équipe est toujours disponible pour les aider à reprendre des activités de recherche comme par exemple les intégrer dans des projets de recherche.

3.2.3 Publications

Les références
- comportant « * » sont acceptées et à paraître.
- indiquées sur fond gris sont les productions phare de chaque catégorie.
- en gras constituent les 20% les meilleures de la production scientifique de chaque catégorie.

Signification des soulignements :
- *simple* : enseignant-chercheur ou chercheur titulaire,
- *double* : doctorant,
- *pointillé* : post-doc, ingénieur ou autre situation.

Articles dans des revues avec comité de lecture et chapitres de livres (ACL)

Articles dans des revues internationales avec comité de lecture sélectif

2022 [RI-21] E. Cartier, A. Onysko, E. Winter-Froemel, E. Zenner, G. Andersen, B. Hilberink-Schulpen, U. Nederstigt, E. Peterson et F. van Meurs. Linguistic repercussions of COVID-19 : A corpus study on four languages. *Open Linguistics*, 8(1) :751–766, 2022.

[RI-20] D. Dessi, F. Osborne, D. Reforgiato Recupero, D. Buscaldi et E. Motta. SCICERO : A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain. *Knowledge-Based Systems*, 258 :109945 (14 pages), 2022.

[RI-19] T. Gaillat, A. Simpkin, N. Ballier, B. Stearns, A. Sousa, M. Bouyé et M. Zarrouk. Predicting CEFR levels in learners of English : The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2) :130–146, 2022.

2021 [**RI-18**] D. Dessi, F. Osborne, D. Reforgiato Recupero, D. Buscaldi et E. Motta. Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain. *Future Generation Computer Systems*, 116 :253–264, 2021.

[**RI-17**] L. Di Rocco, F. Dassereto, M. Bertolotto, D. Buscaldi, B. Catania et G. Guerrini. Sherloc : a knowledge-driven algorithm for geolocating microblog messages at sub-city level. *International Journal of Geographical Information Science*, 35(1) :84–115, 2021.

2020 [RI-16] N. Firoozeh, A. Nazarenko, F. Alizon et B. Daille. Keyword extraction : Issues and methods. *Natural Language Engineering*, 26(3) :259–291, 2020.

[RI-15] A. Grezka. Variabilité et traitement automatique des langues. *Lingvisticae Investigationes*, 43(2) :280–299, 2020.

[RI-14] P. Mogorrón Huerta, A. Grezka et L. Navarro-Brotons. Les variations diatopiques dans les expressions figées. *Lingvisticæ Investigationes*, 43(2) :169–171, 2020.

[RI-13] J. Raad, W. Beek, F. Van Harmelen, J. Wielemaker, N. Pernelle et F. Saïs. Constructing and cleaning identity graphs in the LOD cloud. *Data Intelligence*, 2(3) :323–352, 2020.

[RI-12] D. Symeonidou, V. Armant et N. Pernelle. BECKEY : Understanding, comparing and discovering keys of different semantics in knowledge bases. *Knowledge-Based Systems*, 195 :105708 (26 pages), 2020.

2019 [RI-11] A. Grezka et A. Kijima. L'expression de la perception visuelle : regard franco-japonais. *Journal in English Lexicology (LEXIS)*, 13(13) :19 pages, 2019.

[RI-10] D. Reforgiato Recupero, M. Alam, D. Buscaldi, A. Grezka et F. Tavazoee. Frame-based detection of figurative language in tweets [application notes]. *IEEE Computational Intelligence Magazine*, 14(4) :77–88, 2019.

2018 [RI-9] E. Cartier. Noms propres et innovations lexicales : étude linguistique et statistique à partir de Néoveille. *Cahiers de Lexicologie*, 113(113) :203–224, 2018.

[RI-8] J. Garcia Flores, I. V. Meza Ruiz, E. Colin, C. Gardent, A. Gangemi et L. A. Pineda. Robot experience stories : First person generation of robotic task narratives in $SitLog^1$. *Journal of Intelligent and Fuzzy Systems*, 34(5) :3291–3300, 2018.

2017 [RI-7] M. Alam, D. Reforgiato Recupero, M. Mongiovì, A. Gangemi et P. Ristoski. Event-based knowledge reconciliation using frame embeddings and frame similarity. *Knowledge Based Systems*, 135 :192–203, 2017.

[RI-6] S. Consoli, V. Presutti, D. Reforgiato Recupero, A. G. Nuzzolese, S. Peroni, M. Mongiovì et A. Gangemi. Producing Linked Data for Smart Cities : The Case of Catania. *Big Data Research*, 7 :1–15, 2017.

[RI-5] A. Gangemi, S. Peroni, D. M. Shotton et F. Vitali. The Publishing Workflow Ontology (PWO). *Semantic Web*, 8(5) :703–718, 2017.

[RI-4] A. Gangemi, V. Presutti, D. Reforgiato Recupero, A. G. Nuzzolese, F. Draicchio et M. Mongiovì. Semantic Web Machine Reading with FRED. *Semantic Web*, 8(6) :873–893, 2017.

[RI-3] A. Hernandez-Castañeda, H. Calvo, A. Gelbukh et J. Garcia Flores. Cross-domain deception detection using support vector networks. *Soft Computing*, 21(3) :585–595, 2017.

[RI-2] A. Kemmar, Y. Lebbah, S. Loudni, P. Boizumault et T. Charnois. Prefix-projection global constraint and top-k approach for sequential pattern mining. *Constraints*, 22(2) :265–306, 2017.

[RI-1] A. G. Nuzzolese, V. Presutti, A. Gangemi, S. Peroni et P. Ciancarini. Aemoo : Linked Data exploration based on Knowledge Patterns. *Semantic Web*, 8(1) :87–112, 2017.

Articles dans des revues nationales ou internationales

2022 [RE-7] J. Altmanova, E. Cartier, J. Luzzy, S. Pinto et S. Piscopo. Innovations lexicales dans le domaine de l'environnement et de la biodiversité : le cas de bio en français et en italien. *Neologica, numéro spécial Néologie et environnement*, 16 :85–110, 2022.

[RE-6] A. Borrego, D. Dessi, I. Hernandez, F. Osborne, D. Reforgiato Recupero, D. Ruiz, D. Buscaldi et E. Motta. Completing Scientific Facts in Knowledge Graphs of Research Concepts. *IEEE Access*, 10 :125867–125880, 2022.

2021 [RE-5] E. Cartier et R. Huyghe. La concurrence affixale en diachronie : le cas des préfixes de haut degré en français. *Linx, numéro spécial : Entre vieillissement et innovation : le changement linguistique*, 82 :1–26, 2021.

[RE-4] C. Recanati. La démarche originale de Jean Nicod et son intérêt en sémantique et en sciences cognitives. *Intellectica*, 75(2) :257–284, 2021.

2019 [RE-3] E. Cartier. Neoveille, plateforme de repérage et de suivi des néologismes en corpus dynamique. *Néologica*, 13 :23–54, 2019.

[RE-2] A. Grezka, M. Niziolek et D. Buscaldi. Description de quelques procédés linguistiques de l'ironie, par le biais des tweets sur les transports en commun en français et en polonais. *Studia Romanica Posnaniensia*, 46(1) :43–64, 2019.

2018 [RE-1] A. Grezka, H. Takada et A. Kijima. A record of the lecture given by Dr. Grezka. *Universality and Individuality in Language*, 9 :71–93, 2018. (en japonais).

Chapitres de livre

2021 [CL-11] E. Cartier. *Discourse traditions and computational linguistics*, E. Winter-Froemel et A. S. Octavio de Toledo y Huerta, éditeurs, *Manual of Discourse Traditions in Romance*, volume 30 de *Manuals of Romance Linguistics*, pages 669–690. De Gruyter, 2021.

2020 [CL-10] E. Cartier. *Etude de la variation et du changement lexical à l'ère du numérique*, G. Tallarico, J. Humbley et C. Jacquet-Pfau, éditeurs, *Nouveaux horizons pour la néologie en français*, pages 199–244. Lambert-Lucas, 2020.

[CL-9] E. Cartier. *L'innovation lexicale en français contemporain : étude linguistique, socio-pragmatique et statistique de la presse féminine*, F. Diémoz, G. Dostie, P. Hadermann et F. Lefeuvre, éditeurs, *Le français innovant*, volume 130 de *Sciences pour la communication*, pages 21–56. Peter Lang, 2020.

[CL-8] E. Cartier et J. Lazar. *Anglicismes en français et en tchèque contemporains : le cas des formes en -ing*, R. Mudrochová, E. Cartier et A. Kacprzak, éditeurs, *Les emprunts néologiques et leurs équivalents autochtones, études outillées sur corpus*, volume 2020 de *Philologica*, pages 117–132. Charles University Karolinum Press, 2020.

2019 [CL-7] E. Cartier. *Emprunts en français contemporain : étude linguistique et statistique à partir de la plateforme Néoveille*, A. Kacprzak, R. Mudrochová et J.-F. Sablayrolles, éditeurs, *L'emprunt en question(s)*, pages 145–186. Lambert Lucas, 2019.

2018 [CL-6] D. Legallois, T. Charnois et M. Larjavaara. *Grammar of genres and styles : an overview*, D. Legallois, T. Charnois et M. Larjavaara, éditeurs, *The Grammar of Genres and Styles*, pages 1–14. Numéro 320 Trends in Linguistics. Studies and Monographs [TiLSM]. De Gruyter, 2018.

[CL-5] D. Legallois, T. Charnois et M. Larjavaara. *The balance between quantitative and qualitative literary stylistics : how the method of "motifs" can help*, D. Legallois, T. Charnois et M. Larjavaara, éditeurs, *The Grammar of Genres and Styles*, pages 164–193. Numéro 320 Trends in Linguistics. Studies and Monographs [TiLSM]. De Gruyter, 2018.

[CL-4] C. Roussey, H. Hernandez et H. Zargayouna. *Domain ontologies*, P. A. Fuertes-Olivera, éditeur, *The Routledge Handbook of Lexicography*, Chapitre 14, pages 217–234. Routledge Taylor & Francis group, 2018.

2017 [CL-3] E. Cartier et V. Julie. *Etude de la pénétration des anglicismes de type N ou ADJ(-)Ving à partir d'un corpus contemporain journalistique : les exemples de bashing et shaming en français contemporain*, C. Jacquet-Pfau, A. Napieralski et J.-F. Sablayrolles, éditeurs, *Emprunts néologiques et équivalents autochtones : études interlangues*, pages 11–34. Wydawnictwo Uniwersytetu Łódzkiego, 2017.

[CL-2] B. Daille et A. Nazarenko. *Le tournant des données en traitement automatique des langues*, M. Bouzeghoub et R. Mosseri, éditeurs, *Les big data à découvert*, pages 118–119. CNRS Editions, 2017.

[CL-1] G. Lodi, L. Asprino, A. G. Nuzzolese, V. Presutti, A. Gangemi, D. Reforgiato Recupero, C. Veninata et A. Orsini. *Semantic Web for Cultural Heritage Valorisation*, S. Hai-Jew, éditeur, *Data Analytics in Digital Humanities*, pages 3–37. Springer, 2017.

Communications avec actes (ACT)

Communications dans des conférences internationales avec comité de lecture

2022 [CI-36] Y. Dauxais, U. Zaratiana, M. Laneuville, S. D. Hernandez-Perez, P. Holat et C. Grosman. Towards Automation of Topic Taxonomy Construction. *The 20th International Symposium on Intelligent Data Analysis (IDA 2022), Rennes, France, April 20-22, 2022*, volume 13205 de *Lecture Notes in Computer Science*, pages 26–38, 2022.

[CI-35] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi et E. Motta. CS-KG : A Large-Scale Knowledge Graph of Research Entities and Claims in Computer Science. *The 21st International Semantic Web Conference (ISWC 2022), Hanghzou, China, October 23-27, 2022*, volume 13489 de *Lecture Notes in Computer Science*, pages 678–696, 2022.

[CI-34] G. Felhi, J. Le Roux et D. Seddah. Exploiting Inductive Bias in Transformers for Unsupervised Disentanglement of Syntax and Semantics with VAEs. *The 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL 2022), Seattle, United States, July 10-15, 2022*, pages 5763–5776, 2022.

[CI-33] J. L. Garcia Mendoza, D. Buscaldi et F. Orihuela-Espina. Evaluation of a New Representation for Noise Reduction in Distant Supervision. *The 21st Mexican International Conference on Artificial Intelligence (MICAI 2022), Monterrey, Mexico, October 24-29, 2022*, volume 13613 de *Lecture Notes in Artificial Intelligence*, pages 101–113, 2022.

[CI-32] J.-A. González, D. Buscaldi, E. Sanchis et L.-F. Hurtado. Transformer-Based Models for the Automatic Indexing of Scientific Documents in French. *The 27th International Conference on Applications of Natural Language to Information Systems (NLDB 2022), Valencia, Spain, June 15-17, 2022*, volume 13286 de *Lecture Notes in Computer Science*, pages 60–72. Springer, 2022.

[CI-31] V. Liyanage, D. Buscaldi et A. Nazarenko. A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications. *The 13th Language Resources and Evaluation Conference,(LREC 2022), Marseille, France, June 20-25, 2022*, pages 4692–4700, 2022.

[CI-30] L. Simonne, N. Pernelle et F. Saïs. Counter Effect Rules Mining in Knowledge Graphs. *The 23rd International Conference on Knowledge Engineering and Knowledge Management (EKAW 2022), Bolzano, Italy, September 26-29, 2022*, volume 13514 de *Lecture Notes in Computer Science*, pages 167–173, 2022.

[CI-29] L. Simonne, N. Pernelle, F. Sais et R. Thamopoulos. Discovering Causal Rules in Knowledge Graphs using Graph Embeddings. *IEEE-ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT),*, page 6, Niagara Falls, Canada, 2022.

[CI-28] U. Zaratiana, N. Tomeh, P. Holat et T. Charnois. GNNer : Reducing Overlapping in Span-based NER Using Graph Neural Networks. *The 60th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop (ACL 2022), Dublin, Ireland, May 22-27, 2022*, pages 97–103, 2022.

[**CI-27**] X. Zhang, J. Le Roux et T. Charnois. Higher-Order Dependency Parsing for Arc-Polynomial Score Functions via Gradient-Based Methods and Genetic Algorithm. *The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL/IJCNLP 2022), Online Only, November 20-23, 2022*, pages 1158–1171, 2022.

2021 [CI-26] L. Azevedo, M. d'Aquin, B. Davis et M. Zarrouk. LUX (Linguistic aspects Under eXamination) : Discourse Analysis for Automatic Fake News Classification. *Findings of the Association for Computational Linguistics (ACL/IJCNLP 2021), Bangkok, Thailand, August 1-6, 2021*, pages 41–56, 2021.

[CI-25] J. Lopez Espejel, G. De Chalendar, J. Garcia Flores, T. Charnois et I. V. Meza Ruiz. GeSERA : General-domain Summary Evaluation by Relevance Analysis. *Recent Advances in Natural Language Processing 2021 (RANLP 2021), Varna, Bulgaria, Sep 1–3, 2021*, pages 856–867, 2021.

[CI-24] T. Mecharnia, N. Pernelle, C. Rouveirol, F. Hamdi et L. Chibout. Mining Contextual Rules to Predict Asbestos in Buildings. *The 26th International Conference on Conceptual Structures (ICCS 2021), Berlin, Germany, September 20-22, 2021*, volume 12879 de *Lecture Notes in Artificial Intelligence*, pages 170–184, 2021.

[CI-23] A. Nazarenko, F. Lévy et A. Wyner. A Pragmatic Approach to Semantic Annotation for Search of Legal Texts - An Experiment on GDPR. *The 34th Annual Conference on Legal Knowledge and Information Systems (JURIX 2021), Vilnius, Lithuania, 8-10 December 2021*, volume 346 de *Frontiers in Artificial Intelligence and Applications*, pages 23–32, 2021.

[CI-22] G. Santini, H. Soldano et S. Zevio. Experiments on F-Restricted Bi-pattern Mining. *The 10th International Conference on Complex Networks and Their Applications (COMPLEX NETWORKS X), Madrid, Spain, November 30-December 2, 2021*, volume 1015 de *Studies in Computational Intelligence*, pages 539–550, 2021.

[CI-21] G. Santini, H. Soldano et S. Zevio. Restricted Bi-pattern Mining of Attributed Networks. *The 26th International Conference on Conceptual Structures (ICCS 2021), Virtual Event, September 20-22, 2021*, volume 12879 de *Lecture Notes in Artificial Intelligence*, pages 200–207, 2021.

[CI-20] L. Simonne, N. Pernelle, F. Saïs et R. Thomopoulos. Differential Causal Rules Mining in Knowledge Graphs. *The 11th on Knowledge Capture Conference (K-CAP '21), Virtual Event, USA, December 2-3, 2021*, pages 105–112, 2021.

[CI-19] H. Vornhagen, M. Zarrouk, B. Davis et K. Young. Do city dashboards make sense ? Conceptualising user experiences and challenges in using city dashboards. A case study. *The 22nd Annual International Conference on Digital Government Research (DG.O'21), Omaha, NE, USA, June 9-11, 2021*, pages 219–226, 2021.

[CI-18] X. Zhang, J. Le Roux et T. Charnois. Strength in numbers : averaging and clustering effects in mixture of experts for graph-based dependency parsing. *The 17th International Conference on Parsing Technologies (IWPT 2021), Bangkok, Thailand, August 6, 2021*, pages 106–118, 2021.

2020 [CI-17] E. Cartier. Borrowings Characteristics and Life-Cycles : a Corpus-Driven Study in Contemporary French . *The Fifth International Conference on Language Contact in Times of Globalization (LCTG 5), Klagenfurt, Austria, September 9–10, 2021*, pages 31–32, 2020.

[CI-16] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta et H. Sack. AI-KG : An Automatically Generated Knowledge Graph of Artificial Intelligence. *The 19th International Semantic Web Conference (ISWC 2020), Athens, Greece, November 2-6, 2020*, volume 12507 de *Lecture Notes in Computer Science*, pages 127–143, 2020.

[CI-15] Y. Kankanampati, J. Le Roux, N. Tomeh, D. Taji et N. Habash. Multitask Easy-First Dependency Parsing : Exploiting Complementarities of Different Dependency Representations. *The 28th International Conference on Computational Linguistics (COLING 2020), Barcelona, Spain, December 8-13, 2020*, pages 2497–2508, 2020.

[**CI-14**] A. Khajeh, N. Pernelle, F. Saïs et G. Quercini. Generating Referring Expressions from RDF Knowledge Graphs for Data Linking. *The 19th International Semantic Web Conference (ISWC 2020), Athens, Greece, November 2-6, 2020*, volume 12506 de *Lecture Notes in Computer Science*, pages 311–329, 2020.

2019 [CI-13] L. Di Rocco, D. Buscaldi, M. Bertolotto, B. Catania et G. Guerrini. The role of geographic knowledge in sub-city level geolocation. *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC 2019), Limassol, Cyprus, April 8-12, 2019*, pages 687–689, 2019.

[CI-12] I. Garrido-Marquez, A. Nazarenko et F. Lévy. Combining Free Collaborative and Controlled Annotation in Social Media. *The 6th International Conference on Social Networks Analysis, Management and Security (SNAMS 2019), Granada, Spain, October 22-25, 2019*, pages 572–577, 2019.

[CI-11] J. Le Roux, A. Rozenknop et M. Lacroix. Representation Learning and Dynamic Programming for Arc-Hybrid Parsing. *The 23rd Conference on Computational Natural Language Learning (CoNLL 2019), Hong Kong, China, November 3-4, 2019*, pages 238–248, 2019.

2018 [CI-10] P. Alizadeh, P. Cellier, T. Charnois, B. Crémilleux et A. Zimmermann. An Experimental Approach For Information Extraction in Multi-Party Dialogue Discourse. *The19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2018), Hanoi, Vietnam, March 18-24, 2018, Revised Selected Papers*, volume 13396 de *Lecture Notes in Computer Science*, pages 180–192, 2018. http ://www.cicling.org/2018/intranet/pre-print/papers/paper_173.pdf.

[CI-9] I. Garrido Marquez, J. Garcia Flores, F. Lévy et A. Nazarenko. Controlling the drift of semantic indexing systems. *The 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD 2018), Seville, Spain, September 18-20, 2018*, pages 199–206, 2018.

[CI-8] A. Nazarenko, F. Lévy et A. Z. Wyner. An Annotation Language for Semantic Search of Legal Sources. *The 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7-12, 2018*, pages 1096–1100, 2018.

2017 [CI-7] H. Calvo, A. Hernandez-Castaneda et J. Garcia Flores. Author Identification Using Latent Dirichlet Allocation. *The 18th Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017), Budapest, Hungary, April 17-23, 2017, Revised Selected Papers*, volume 10762 de *Lecture Notes in Computer Science*, page 12, 2017.

[CI-6] E. Cartier. Néoveille, a Web Platform for Neologism Tracking. *The 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), Valencia, Spain, April 3-7, 2017, Software DemonstrationsT*, pages 95–98, 2017.

[CI-5] C. Corro, J. Le Roux et M. Lacroix. Efficient Discontinuous Phrase-Structure Parsing via the Generalized Maximum Spanning Arborescence. *The 2017 Conference on Empirical Methods in Natural Language Processing (EMLNP 2017), Copenhagen, Denmark, September 7–11, 2017*, pages 1644 – 1654, 2017.

[**CI-4**] K. Gábor, H. Zargayouna, I. Tellier, D. Buscaldi et T. Charnois. Exploring Vector Spaces for Semantic Relations. *The 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), Copenhagen, Denmark, September 9-11, 2017*, pages 1814–1823, 2017.

[CI-3] A. Meloni, D. Reforgiato Recupero et A. Gangemi. AMR2FRED, A Tool for Translating Abstract Meaning Representation to Motif-Based Linguistic Knowledge Graphs. *The 14th Extended Semantic Web Conference (ESWC 2017) Satellite Events, Portoro\vz, Slovenia, May 28 - June 1, 2017, Revised Selected Papers*, volume 10577 de *Lecture Notes in Computer Science*, pages 43–47, 2017.

[CI-2] J. Reynaud, M. Alam, Y. Toussaint et A. Napoli. A Proposal for Classifying the Content of the Web of Data Based on FCA and Pattern Structures. *23rd International Symposium on Methodologies for Intelligent Systems (ISMIS 2017), Warsaw, Poland, June 26-29, 2017*, volume 10352 de *Lecture Notes in Artificial Intelligence*, pages 684–694, 2017.

[CI-1] A. Wyner, F. Gough, F. Lévy, M. Lynch et A. Nazarenko. On annotation of the textual contents of scottish legal instruments. *The 13th Annual Conference Legal Knowledge and Information Systems (Jurix 2017), Luxembourg, December 13-15, 2017*, volume 302 de *Frontiers in Artificial Intelligence and Applications*, pages 101–106, 2017.

Autres articles dans des conférences ou workshops à comité de lecture

2023 [CO-64] E. Cartier. Semantic Change Tracking Through the Prism of Distributionnalism and Construction Grammars : an Experiment in Contemporary French. *The 14th International Cognitive Linguistics Conference : Linguistic Diversity and Cognitive Linguistics (ICLC-14), 10—14 July 2017 Tartu, Estonia*, pages 214–215, 2023.

[CO-63] A. Khajeh, N. Pernelle et F. Saïs. Enrichissement de règles de Horn par des predicats numériques. *Extraction et Gestion des Connaissances (EGC 2023), Lyon, France, 16 - 20 janvier 2023*, volume E-39 de *Revue des Nouvelles Technologies de l'Information*, pages 425–432, 2023.

2022 [CO-62] N.-M. Boukhatem, D. Buscaldi et L. Liberti. Empirical Comparison of Semantic Similarity Measures for Technical Question Answering. *The 26th European Conference on Advances in Databases and Information Systems : New Trends in Database and Information Systems (ADBIS 2022) - Shorts Papers, Turin, Italy, September 5-8, 2022*, volume 1652 de *Communications in Computer and Information Science*, pages 167–177, 2022. (Short paper).

[CO-61] E. Cartier. Diachronic Semantic Evolution Automatic Tracking : a pilot study in Modern and Contemporary French combining Dependency Analysis and Contextual Embeddings. *The 4th Globalex Workshop on Lexicography and Neology (GWLN4 @EURALEX22), Mannheim, Germany, July 15, 2022*, 2022. https ://videolectures.net/euralex2022_cartier_diachronic_semantic/.

[CO-60] N. Elkhbir, N. Tomeh et T. Charnois. ArabIE : Joint Entity, Relation and Event Extraction for Arabic. *The 7th Arabic Natural Language Processing Workshop (WANLP@EMNLP 2022), Abu Dhabi, United Arab Emirates, December 8, 2022*, pages 331–345, 2022.

[CO-59] G. Felhi, J. Le Roux et D. Seddah. Towards Unsupervised Content Disentanglement in Sentence Representations via Syntactic Roles. *The 1st Controllable Generative Modeling in Language and Vision Workshop (CtrlGen Workshop) at NeurIPS 2021, Virtual Conference, December 13, 2021*, virtual, France, 2022.

[**CO-58**] M. Kamal Eddine, N. Tomeh, N. Habash, J. Le Roux et M. Vazirgiannis. AraBART : a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization. *The Seventh Arabic Natural Language Processing Workshop (WANLP@EMNLP 2022), Abu Dhabi, United Arab Emirates, December 8, 2022*, pages 31–42, 2022.

[CO-57] A. Khajeh, N. Pernelle, F. Saïs et G. Quercini. Knowledge Graph Refinement based on Triplet BERT-Networks. *Deep Learning meets Ontologies and Natural Language Processing (DeepOntoNLP), 3rd International Workshop, in conjunction with ESWC 2022, Hersonissos, Greece, May 29-June 2, 2022*, pages 1–9, The 3rd International Workshop on Deep Learning meets Ontologies, 2022.

[CO-56] L. Simonne, N. Pernelle et F. Saïs. Découverte de règles causales dans les graphes de connaissances à l'aide de plongements dans les graphes. *33es Journées francophones d'Ingénierie des Connaissances (IC 2022), Saint-Étienne, France, 29 juin–1 juillet 2022*, pages 1–10, 2022.

[CO-55] U. Zaratiana, N. Elkhbir, P. Holat, N. Tomeh et T. Charnois. Global Span Selection for Named Entity Recognition. *The 2022 Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS@EMNLP 2022), Abu Dhabi, United Arab Emirates, December 7, 2022*, pages 11–17, 2022.

[**CO-54**] U. Zaratiana, N. Elkhbir, D. Núñez, P. Holat, N. Tomeh et T. Charnois. DyREx : Dynamic Query Representation for Extractive Question Answering. *The 2nd Workshop on Efficient Natural Language and Speech Processing (ENSLP@NeurIPS 2022), New Orleans, LO, United States, December 2, 2022*, pages 1–6. 2nd Workshop on Efficient Natural Language and Speech Processing @ Neurips 2022, 2022.

[CO-53] U. Zaratiana, N. Tomeh, P. Holat et T. Charnois. Named Entity Recognition as Structured Span Prediction. *The First Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS@EMNLP 2022), Abu Dhabi, United Arab Emirates, December 7-8, 2022*, pages 1–10, 2022.

2021 [CO-52] G. Felhi, J. Le Roux et D. Seddah. Challenging the Semi-Supervised VAE Framework for Text Classification. *The Second Workshop on Insights from Negative Results in NLP, Punta Cana, Dominican Republic, November 10, 2021*, pages 136–143, 2021.

[CO-51] F. Lovera, Y. Cardinale, D. Buscaldi, T. Charnois et N. H. Masun. Deep Learning Enhanced with Graph Knowledge for Sentiment Analysis. *6th International Workshop on Explainable Sentiment Mining and Emotion Detection (X-SENTIMENT 2021) co-located with 18th Extended Semantic Web Conference 2021 (ESWC 2021), Hersonissos, Greece, June 6 –7, 2021*, volume 2918 de *CEUR Workshop Proceedings*, pages 74–86, 2021.

[CO-50] T. Mecharnia, L. Chibout Khelifa, F. Hamdi, N. Pernelle et C. Rouveirol. Découverte de règles contextuelles pour prédire la présence d'amiante dans les bâtiments. *32èmes Journées francophones d'Ingénierie des Connaissances (IC 2021), Bordeaux, France, June 30 - July 2, 2021*, pages 81–88, 2021.

[CO-49] L. Simonne, N. Pernelle et F. Saïs. Fouille de règles différentielles causales dans les graphes de connaissances. *Extraction et Gestion des Connaissances (EGC 2021), 25-29 Janvier 2021, Montpellier, France*, volume E-37 de *Revue des Nouvelles Technologies de l'Information*, pages 293–300, 2021.

2020 [CO-48] P. Alizadeh, I. Vladimir et J. Garcia Flores. Apprentissage par renforcement pour la recherche d'experts sur le web. *Extraction et Gestion des Connaissances (EGC 2020), Brussels, Belgium, January 27-31, 2020*, volume E-36 de *Revue des Nouvelles Technologies de l'Information*, pages 261–268. Éditions RNTI, 2020.

[CO-47] D. Buscaldi, G. Felhi, D. Ghoul, J. Le Roux, G. Lejeune et X. Zhang. Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition), Atelier DÉfi Fouille de Textes, Nancy, France, June 8-19, 2020*, pages 14–25, 2020.

[CO-46] T. Gaillat, N. Ballier, A. Sousa, M. Bouyé, A. Simpkin, B. Stearns et M. Zarrouk. Un prototype en ligne pour la prédiction du niveau de compétence en anglais des productions écrites. *6ème conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition), Nancy, France, June 8-19, 2020*, volume 4, pages 30–33, 2020.

[CO-45] B. Ghanem, D. Buscaldi et P. Rosso. TexTrolls : Identifying Trolls on Twitter with Textual and Affective Features. *The 2020 Workshop on Online Misinformation and Harm-Aware Recommender Systems (OHARS'20) co-located with 14th ACM Conference on Recommender Systems (RecSys 2020), Rio de Janeiro, Brazil, September 25, 2020*, volume 2758 de *CEUR Workshop Proceedings*, pages 4–22, 2020.

[CO-44] A. Grezka. Verbes de perception et traitement de la polysémie : pourquoi et comment ? *Actes du colloque international "La perception : langue, discours, cognition", CELIS0, Paris, France, 6-7 décembre 2019*, pages 29–44. Irina Thomières-Shakovskaya, 2020.

[CO-43] A. Khajeh, N. Pernelle et F. Saïs. Découverte d'expressions référentielles dans les graphes de connaissances. *Extraction et Gestion des Connaissances (EGC 2020), Brussels, Belgium, January 27-31, 2020*, volume E-36 de *Revue des Nouvelles Technologies de l'Information*, pages 357–364, 2020.

[CO-42] A. Khajeh, N. Pernelle, F. Saïs et G. Quercini. RE-miner for data linking results for OAEI 2020. *The 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (OM@ISWC 2020), Athens, Greece, November 2, 2020*, volume 2788 de *CEUR Workshop Proceedings*, pages 211–215, 2020.

[CO-41] A. Sousa, N. Ballier, T. Gaillat, B. Stearns, M. Zarrouk, A. Simpkin et M. Bouyé. From Linguistic Research Projects to Language Technology Platforms : A Case Study in Learner Data. *Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020), Marseille, France, May 11-12, 2020*, Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020), pages 112–120, 2020.

[CO-40] A. Sousa, M. d'Aquin, M. Zarrouk et J. Holloway. Person-Independent Multimodal Emotion Detection for Children with High-Functioning Autism. *The First Workshop on Artificial Intelligence for Function, Disability, and Health co-located with the 2020 International Joint Conference on Artificial Intelligence - Pacific Rim Conference on Artificial Intelligence (IJCAI-PRICAI 2020), Online, January 7-8, 2021*, volume 2760 de *CEUR Workshop Proceedings*, pages 14–20, 2020.

[CO-39] S. Zevio, G. Santini, H. Soldano, H. Zargayouna et T. Charnois. A Combination of Semantic Annotation and Graph Mining for Expert Finding in Scholarly Data. *The 2020 Workshop on Graph Embedding and Mining Worshop (GEM 2020) co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2020), Virtual Event, September 18, 2020*, page 16, 2020.

2019 [CO-38] D. Buscaldi, D. Dessì, E. Motta, F. Osborne et D. Reforgiato Recupero. Mining Scholarly Data for Fine-Grained Knowledge Graph Construction. *The Workshop on Deep Learning for Knowledge Graphs (DL4KG2019) Co-located with the 16th Extended Semantic Web Conference 2019 (ESWC 2019), Portoroz, Slovenia, June 2, 2019*, volume 2377 de *CEUR Workshop Proceedings*, pages 21–30, 2019.

[CO-37] D. Buscaldi, D. Dessì, E. Motta, F. Osborne et D. Reforgiato Recupero. Mining Scholarly Publications for Scientific Knowledge Graph Construction. *The 16th Extended Semantic Web Conference - Satellite Events (ESWC 2019 Satellite Events), Portorož, Slovenia, June 2-6, 2019, Revised Selected Papers*, volume 11762 de *Lecture Notes in Computer Science*, pages 8–12, 2019.

[CO-36] D. Buscaldi, D. Ghoul, J. Le Roux et G. Lejeune. Indexation et appariements de documents cliniques pour le Deft 2019. *Edition 2019 du Défi Fouille de Textes (DEFT 2019), Toulouse, France, 9-15 Mai 2019*, pages 49–56, 2019.

[CO-35] A. Grezka, M. Niziolek et D. Buscaldi. Description linguistique et détection automatique de l'ironie dans les tweets en français et en polonais. *5ème Forum franco-polonais de la recherche et de l'innovation, Varsovie, Pologne, 19 Novembre 2019*. 5ème forum franco-polonais de la recherche et de l'innovation, 2019. (poster).

[CO-34] H. Murrieta, I. Meza, P. Alizadeh et J. Garcia Flores. Towards Identifying for Evidence of Drain Brain from Web Search Results using Reinforcement Learning. *LatinX in AI*

Research Workshop (LXAI) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, December 8-14, 2019, pages 1–3, 2019. (poster).

[CO-33] S. Zevio, G. Santini, H. Zargayouna et T. Charnois. Fouille de texte et fouille de graphe appliquées à la recherche d'experts. *30èmes journées francophones d'Ingénierie des Connaissances (IC 2019), Toulouse, France, 2-4 juillet 2019*, pages 222–223, 2019.

2018 [CO-32] P. Alizadeh, P. Cellier, T. Charnois, B. Crémilleux et A. Zimmermann. Etude expérimentale d'extraction d'information dans des retranscriptions de réunions. *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2018), Rennes, France, May 14-18, 2018*, pages 465–472, rennes, france, 2018.

[CO-31] D. Buscaldi. Improving access to scientific literature : a semantic IR perspective. *The 5th Spanish Conference on Information Retrieval (CERI 2018), Zaragoza, Spain, June 26-27, 2018*, page 1 :1, 2018.

[CO-30] D. Buscaldi. Tweetaneuse @ AMI EVALITA2018 : Character-based Models for the Automatic Misogyny Identification Task (Short Paper). *The Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 de *CEUR Workshop Proceedings*, 2018.

[CO-29] D. Buscaldi, J. Le Roux et G. Lejeune. Modèles en Caractères pour la Détection de Polarité dans les Tweets. *13ème Edition de l'Atelier Défi Fouille de Textes (DEFT@CORIA/TALN 2018), Rennes, France, 15 mai 2018*, pages 249–258, 2018.

[CO-28] E. Cartier. Néologie sémantique : modélisation, expérimentations automatiques multilingues dans le cadre de Néoveille. *4ème Congrès international de néologie des langues romanes (CINEO 2018), Lyon, France, July 4-6, 2018*, pages 1–2, 2018.

[CO-27] E. Cartier. Néonaute : mining web archives for linguistic analysis. *IIPC Web archiving Conference 2018, Wellington (New-Zealand), November 12-15, 2018.*, 2018. sans proceedings.

[CO-26] E. Cartier. Néoveille, an Automatic System for Lexical Units Life-Cycle Tracking. *The XVIII EURALEX International Congress, Ljubljana, Slovenia, July 17-21, 2018*, pages 34–35, XVIII EURALEX International Congress, Ljubljana, Slovenia, 2018. sans proceedings.

[CO-25] E. Cartier, L. Galand, P. Stirling et S. Aubry. Néonaute, Enrichissement sémantique pour la recherche d'information. *25ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2018), Rennes, France, May 14-18, 2018*, pages 201–204, 2018.

[CO-24] E. Cartier, J.-F. Sablayrolles, N. Boutmgharine, J. Humbley, M. Bertocci, C. Jacquet-Pfau, N. Kubler et G. Tallarico. Détection automatique, description linguistique et suivi des néologismes en corpus : point d'étape sur les tendances du français contemporain. *6ème Congrès Mondial de Linguistique Française, Mons, Belgique, 9-13 juillet 2018*, volume 46 de *SHS Web of Conferences*, pages 08002 : 1–20, 2018.

[**CO-23**] K. Gábor, D. Buscaldi, A.-K. Schumann, B. Q. Zadeh, H. Zargayouna et T. Charnois. SemEval-2018 Task 7 : Semantic Relation Extraction and Classification in Scientific Papers. *The 12th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2018), New Orleans, Louisiana, USA, June 5-6, 2018*, pages 679–688, 2018.

[CO-22] K. Gábor, N. Lechevrel, I. Tellier, D. Buscaldi, H. Zargayouna et T. Charnois. Apport des dépendances syntaxiques et des patrons séquentiels à l'extraction de relations. *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2018) , Rennes, France, May 14-18, 2018*, pages 557–566, 2018.

[CO-21] S. Zevio. Identifying Transmedia Works from User-Generated Knowledge Bases : Japanese Pop Culture Study Case. *The 8th Joint International Semantic Technology Conference co-located with the 8th Joint International Semantic Technology Conference (JIST 2018), Awaji City, Hyogo, Japan, November 26-28, 2018 (Workshops and Posters)*, volume 2293 de *CEUR Workshop Proceedings*, pages 33–42, 2018.

[CO-20] S. Zevio. Knowledge Discovery and Enrichment from Scholarly Data for Expert Finding. *The EKAW Doctoral Consortium 2018 co-located with the 21st International Conference on Knowledge Engineering and Knowledge Management Doctoral Consortium (EKAW 2018), , Nancy, France, November 13, 2018*, volume 2306 de *CEUR Workshop Proceedings*, pages 1–8, 2018.

[CO-19] S. Zevio, H. Zargayouna, G. Santini et T. Charnois. Vers une cartographie automatique des thématiques et profils d'experts associés à une conférence scientifique : 9 ans d'ateliers recherche d'information semantique. *10ème Atelier Recherche d'Information Sémantique (RISE@CORIA/TALN 2018), Rennes, France, 15 Mai 2018*, rennes, france, 2018.

2017 [CO-18] M. Alam, D. Reforgiato Recupero, M. Mongiovì, A. Gangemi et P. Ristoski. Frame Embeddings for Event-Based Knowledge Reconciliation. *The ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23-25, 2017*, volume 1963 de *CEUR Workshop Proceedings*, 2017.

[CO-17] M. Alam, D. Reforgiato Recupero, M. Mongiovì, A. Gangemi et P. Ristoski. Reconciling Event-Based Knowledge Through RDF2VEC. *The International Workshops on Hybrid Statistical Semantic Understanding and Emerging Semantics, and Semantic Statistics co-located with 16th International Semantic Web Conference (HybridSemStats@ISWC 2017), Vienna, Austria, October 22, 2017*, volume 1923 de *CEUR Workshop Proceedings*, pages Article 3 : 1–10, 2017.

[CO-16] I. Bannour, H. Zargayouna et A. Nazarenko. Modèle de recherche d'information sémantique en graphe : interrogation par propagation d'activation. *28èmes Journées francophones d'Ingénierie des Connaissances (IC 2017), Caen, France, July 3-7, 2017*, pages 181–192, 2017.

[CO-15] D. Buscaldi, A. Grezka et G. Lejeune. Tweetaneuse : Fouille de motifs en caractères et plongement lexical à l'assaut du DEFT 2017. *13ème Défi Fouille de Texte (DEFT 2017), TALN 2017, Orléans, France, 26-30 Juin, 2017*, pages 65–76, 2017.

[CO-14] D. Buscaldi, S. D. Hernandez-Perez et T. Charnois. Classification of keyphrases from scientific publications using WordNet and word embeddings. *1er Atelier Valorisation et Analyse des Données de la Recherche (VADOR 2017) durant la 35e édition du congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2017), Toulouse, France, May 31, 2017*, volume 1860 de *CEUR Workshop Proceedings*, pages 50–57, 2017.

[CO-13] D. Buscaldi et B. Priego. LIPN-UAM at EmoInt-2017 : Combination of Lexicon-based features and Sentence-level Vector Representations for Emotion Intensity Determination. *The 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA@EMNLP 2017), Copenhagen, Denmark, September 8, 2017*, pages 255–258, 2017.

[CO-12] E. Cartier. Detection and Life-cycles Tracking of Neologisms : Experiments in the Neoveille Project. *The Dynamics of Lexical Innovation Workshop 2017, Munich, Germany, June 28-30, 2017*, pages 8–8, 2017.

[CO-11] E. Cartier. Sémantique lexicale et distributionnalisme : éléments pour le repérage automatique du sens en corpus. *7ème Colloque international sur la Représentations du Sens Linguistique (RSL VII) Sherbrooke, Québec, Canada, 25-27 octobre 2017*, 2017.

[CO-10] E. Cartier, G. Lejeune, K. Gábor et T. Charnois. Neoveille, a System for Multilingual Online Neologism Tracking. *The 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2017), Budapest, Hungary, April 17–23, 2017*, pages 13–25, 2017. (Supplementary Proceedings).

[CO-9] C. Corro et J. Le Roux. Transforming Dependency Structures to LTAG Derivation Trees. *The 13th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+13), Umea, Sweden, September 4–6, 2017*, pages 112–121, 2017.

[CO-8] I. Garrido Marquez, A. Nazarenko, F. Lévy et J. Garcia Flores. Mesurer la qualité des systèmes de catégories de blogs. *28èmes Journées francophones d'Ingénierie des Connaissances (IC 2017), Caen, France, July 3-7, 2017*, pages 217–222, 2017.

[CO-7] S. D. Hernandez-Perez, D. Buscaldi et T. Charnois. LIPN at SemEval-2017 Task 10 : Filtering Candidate Keyphrases from Scientific Publications with Part-of-Speech Tag Sequences to Train a Sequence Labeling Model. *The 11th International Workshop on Semantic Evaluation (SemEval@ACL 2017), Vancouver, Canada, August 3-4, 2017*, pages 995–999, 2017.

[CO-6] N. Lechevrel, K. Gábor, I. Tellier, T. Charnois, H. Zargayouna et D. Buscaldi. Combining syntactic and sequential patterns for unsupervised semantic relation extraction. *The*

2017 Workshop on Interactions between Data Mining and Natural Language Processing 2017 (DMNLP 2017) co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2017) Skopje, Macedonia, September 22, 2017, volume 1880 de *CEUR Workshop Proceedings*, pages 81–84, 2017.

[CO-5] G. Lejeune et E. Cartier. Character Based Pattern Mining for Neology Detection. *The First Workshop on Subword and Character Level Models in NLP, Copenhagen, Denmark, September 7, 2017*, pages 25–30, 2017.

[**CO-4**] A. Meroño Peñuela, R. Hoekstra, A. Gangemi, P. Bloem, R. de Valk, B. Stringer, B. Janssen, V. de Boer, A. Allik, S. Schlobach et K. R. Page. The MIDI Linked Data Cloud. *The 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 21-25, 2017*, volume 10588 de *Lecture Notes in Computer Science*, pages 156–164, 2017.

[CO-3] N. Mimouni, A. Nazarenko et S. Salotti. Answering Complex Queries on Legal Networks : A Direct and a Structured IR Approaches. *AI Approaches to the Complexity of Legal Systems - AICOL International Workshops 2015-2017 : AICOL-VI@JURIX 2015, AICOL-VII@EKAW 2016, AICOL-VIII@JURIX 2016, AICOL-IX@ICAIL 2017, and AICOL-X@JURIX 2017, Revised Selected Papers*, volume 10791 de *Lecture Notes in Computer Science*, pages 451–464, 2017.

[CO-2] C. Santos, A. Gangemi et M. Alam. Detecting and Editing Privacy Policy Pitfalls on the Web. *The 1st Workshop on Technologies for Regulatory Compliance co-located with the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017), Luxembourg, December 13, 2017*, volume 2049 de *CEUR Workshop Proceedings*, pages 87–99, 2017.

[CO-1] H. Zargayouna, C. Roussey et S. Ouardani. Annotation sémantique à partir de textes : Cas des observations dans les Bulletins de Santé du végétal. *9ème atelier Recherche d'Information SEmantique (RISE 2017), Caen, France, 4 Juillet 2017*, pages 1–8, 2017.

Direction d'ouvrages (DO)

Livres

2018 [LI-1] D. Legallois, T. Charnois et M. Larjavaara. *The Grammar of genres and styles. From Discrete to Non-Discrete Units*, volume 320 de *Trends in Lingustics book series*. De Gruyter Mouton, 2018. to appear.

Édition d'ouvrages collectifs

2023 [ED-17] A. Kacprzak, E. Cartier et R. Mudrochová. *La néologie par emprunt en français, en polonais et en tchèque : tendances actuelles*, volume 58 de *Études de linguistique, Littérature et Art*. Peter Lang, 2023.

2022 [ED-16] M. Alam, D. Buscaldi, M. Cochez, F. Osborne, D. Reforgiato Recupero et H. Sack. *Special Issue on Deep Learning and Knowledge Graphs*, volume 13 de *Semantic Web*. AKA, 2022.

[**ED-15**] H. Bouamor, H. S. Al-Khalifa, K. Darwish, O. Rambow, F. Bougares, A. Abdelali, N. Tomeh, S. Khalifa et W. Zaghouani. *The 7th Arabic Natural Language Processing Workshop (WANLP@EMNLP 2022), Abu Dhabi, United Arab Emirates, December 8, 2022*. Association for Computational Linguistics, 2022.

[ED-14] D. Buscaldi, M. Dragoni, F. Frasincar et D. Reforgiato Recupero. *Special Issue on Sentimental Health : Future Generation Sentiment Analysis Systems*, volume 129 de *Future Generation Computer Systems*. Elsevier, 2022.

[ED-13] B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali et P. Buitelaar. *The 2nd Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI 2022), Dublin, Ireland, May 27, 2022*. Association for Computational Linguistics, 2022.

[ED-12] V. Khurshudian, N. Tomeh, D. Nouvel, A. Donabédian et C. Vidal-Gorène. *The LREC 2022 Workshop on Processing Language Variation : Digital Armenian (DigitAm), Marseille, France, June 20, 2022*. European Language Resources Association (ELRA), 2022.

[ED-11] B. Piwowarski, F. Boudin, G. Dias, J.-P. Chevallet, J. Moreno, M. Roche et T. Charnois. *Actes de la 3ème journée : Technologies du langage humain et accès interactif à l'information, Paris, Avril 2022*. 2022.

2021 [ED-10] S. Ben Abbès, R. Hantach, P. Calvez, D. Buscaldi, D. Dessì, M. Dragoni, D. Reforgiato Recupero et H. Sack. *The 2nd International Workshop on Deep Learning meets Ontologies and Natural Language Processing (DeepOntoNLP 2021) & 6th International Workshop on Explainable Sentiment Mining and Emotion Detection (X-SENTIMENT 2021) co-located with 18th Extended Semantic Web Conference (ESWC 2021), Hersonissos, Greece, June 6-7, 2021*, volume 2918 de *CEUR Workshop Proceedings*. 2021.

[ED-9] B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali et P. Buitelaar. *The First Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI@EACL 2021), Online, April 19, 2021*, *The First Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI@EACL 2021), Online, April 19, 2021*. Association for Computational Linguistics, 2021.

[ED-8] N. Habash, H. Bouamor, H. M. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha et S. Touileb. *The 6th Arabic Natural Language Processing Workshop, (WANLP 2021), Kyiv, Ukraine, April 9, 2021*. Association for Computational Linguistics, 2021.

2020 [ED-7] M. Alam, D. Buscaldi, M. Cochez, F. Osborne, D. Reforgiato Recupero et H. Sack. *The 2020 Workshop on Deep Learning for Knowledge Graphs (DL4KG2020) co-located with the 17th Extended Semantic Web Conference 2020 (ESWC 2020), Heraklion, Greece, June 2, 2020*, volume 2635 de *CEUR Workshop Proceedings*. 2020.

[ED-6] R. Mudrochová, E. Cartier et A. Kaprczak. *Les emprunts néologiques et leurs équivalents autochtones, études outillées sur corpus*, volume 2020 de *Philologica*. Charles University Karolinum Press, 2020.

[ED-5] I. Zitouni, M. Abdul-Mageed, H. Bouamor, F. Bougares, M. El-Haj, N. Tomeh et W. Zaghouani. *The 5th Arabic Natural Language Processing Workshop (WANLP@COLING 2020), Barcelona, Spain, December 12, 2020*. Association for Computational Linguistics, 2020.

2018 [ED-4] D. Buscaldi, A. Gangemi et D. Reforgiato Recupero. *The 5th Semantic Web Challenges (SemWebEval 2018) at ESWC 2018, Heraklion, Greece, June 3-7, 2018, Revised Selected Papers*, volume 927 de *Communications in Computer and Information Science*. Springer, 2018.

[ED-3] D. Reforgiato Recupero, M. Dragoni, D. Buscaldi, M. Alam et E. Cambria. *The 4th Workshop on Sentic Computing, Sentiment Analysis, Opinion Mining, and Emotion Detection (EMSASW 2018) Co-located with the 15th Extended Semantic Web Conference 2018 (ESWC 2018), Heraklion, Greece, June 4, 2018*, volume 2111 de *CEUR Workshop Proceedings*. 2018.

2017 [ED-2] P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens et Y. Toussaint. *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing 2017 (DMNLP 2017) co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2017) Skopje, Macedonia, September 22, 2017*, volume 880 de *CEUR Workshop Proceedings*. 2017.

[ED-1] N. Habash, M. T. Diab, K. Darwish, W. El-Hajj, H. S. Al-Khalifa, H. Bouamor, N. Tomeh et M. El-Haj. *The 3rd Arabic Natural Language Processing Workshop, (WANLP 2017@EACL), Valencia, Spain, April 3, 2017*. Association for Computational Linguistics, 2017.

Logiciels et Ressources (LO)

2022 [LO-27] D. Antonov et D. Buscaldi. Relation-prediction-from-embeddings. https://github.com/chejuro/Relation-prediction-from-embeddings, 2022.

[LO-26] D. Buscaldi, D. Dessì, D. Reforgiato Recupero, F. Osborne et E. Motta. CS-KG : A Large-Scale Knowledge Graph of Research Entities and Claims in Computer Science. http://scholkg.kmi.open.ac.uk/, 2022.

[LO-25] M. Daychman et D. Buscaldi. Prédiction de relations AIKG / Wordnet (sémantique). https://github.com/Pylyr/Internship_SBERT, 2022.

[LO-24] G. Felhi, J. Le Roux et D. Seddah. QKVAE : A model for Unsupervised Disentanglement of Syntax and Semantics (code). https://github.com/ghazi-f/QKVAE, 2022.

[LO-23] J. A. Gonzalez, D. Buscaldi, L. Hurtado et E. Sanchis. MBartCopyGenerator. https://gist.github.com/jogonba2/, 2022.

[LO-22] M. Kamal Eddine, N. Tomeh, J. Le Roux, N. Habash et M. Vazirgiannis. AraBART : a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization (code). https://huggingface.co/moussaKam/AraBART, 2022.

[LO-21] U. Zaratiana, N. Elkhbir, P. Holat, N. Tomeh et T. Charnois. Global Span Selection for Named Entity Recognition (code). https://github.com/urchade/global-span-selection, 2022.

[LO-20] U. Zaratiana, N. Elkhbir, D. H. Nunez Fernandez, P. Holat, N. Tomeh et T. Charnois. DyREx : Dynamic Query Representation for Extractive Question Answering (code). https://github.com/urchade/DyReX, 2022.

[LO-19] U. Zaratiana, P. Holat, N. Tomeh et T. Charnois. Hierarchical Transformer Model for Scientific Named Entity Recognition (code). https://github.com/urchade/HNER, 2022.

[LO-18] U. Zaratiana, N. Tomeh, P. Holat et T. Charnois. GNNer : Reducing Overlapping in Span-based NER Using Graph Neural Networks (code). https://github.com/urchade/gnner, 2022.

[LO-17] U. Zaratiana, N. Tomeh, P. Holat et T. Charnois. Named Entity Recognition as Structured Span Prediction (code). https://github.com/urchade/span-structured-prediction, 2022.

[LO-16] X. Zhang, J. Le Roux et T. Charnois. Basic Parser, linear and second order. https://github.com/kidlestar/BParser, 2022.

[LO-15] X. Zhang, J. Le Roux et T. Charnois. Higher-Order Dependency Parsing for Arc-Polynomial Score Functions via Gradient-Based Methods and Genetic Algorithm (code). https://github.com/kidlestar/PolyParser, 2022.

2021 [LO-14] D. Buscaldi. Sentence Representations based on Dependency Graphs and Graph Convolutional Networks (DG-GCN). https://github.com/dbuscaldi/PreTens22, 2021.

[LO-13] A. Grezka, J. Garcia Flores et T. Charnois. ChêneTAL : plate-forme d'intégration d'outils de traitement automatique des langues. *XXIVe École Doctorale des Pays de Vysegràd, Autour du texte dans les études françaises„ Cracovie, Pologne 23-25 septembre 2021*, 2021.

[LO-12] J. Lopez Espejel, P. Alizadeh et J. Garcia Flores. RENFO / DiasporasRL (code). https://github.com/jorgegarciaflo/DiasporasRL, 2021.

[LO-11] J. Lopez Espejel, G. De Chalendar, J. Garcia Flores, T. Charnois et I. V. Meza Ruiz. GeSERA : General-domain Summary Evaluation by Relevance Analysis (code). https://github.com/JessicaLopezEspejel/GeSERA, 2021.

[LO-10] X. Zhang, J. Le Roux et T. Charnois. Strength in numbers : averaging and clustering effects in mixture of experts for graph-based dependency parsing (code). https://github.com/kidlestar/MOE, 2021.

2020 [LO-9] Y. Kankanampati, J. Le Roux, N. Tomeh, D. Taji et N. Habash. Multitask Easy-First Dependency Parsing : Exploiting Complementarities of Different Dependency Representations (code). https://github.com/yash-reddy/MEF_parser, 2020.

2019 [LO-8] E. Cartier. Neoveille, plateforme de repérage et de suivi des néologismes en corpus dynamique (outil en ligne). https://tal.lipn.univ-paris13.fr/neoveille/, 2019.

[LO-7] J. Garcia Flores, H. Ferreira, A. Okotan, F. Mantilla et F. Abdoulvahide. Pamparios : Traducteur espagnol - Wixárika (langue amérindienne) (démo). https://tal.lipn.univ-paris13.fr/wixarika/, 2019.

[LO-6] A. Grezka. Morfetik, ressource morphologique. https://tal.lipn.univ-paris13.fr/morfetik, 2019.

[LO-5] H. Kaddouh, M. Labiodh, N. Ghourabi, J. Garcia Flores, A. Hafsa et C. Ourtirane. YuCorrectMaya : validation app of comparable corpus for the Maya Yucatec language (code). https://github.com/rcln/maya, 2019.

2018 [LO-4] D. Buscaldi. Tweetaneuse @ AMI EVALITA2018 : Character-based Models for the Automatic Misogyny Identification Task (code). https://github.com/dbuscaldi/tweetaneuse2018, 2018.

[LO-3] I. Garrido Marquez, J. Garcia Flores, F. Lévy et A. Nazarenko. Kilroy : tag prediction and semantic indexing system (code). https://github.com/IvanGarridoMarquez/predictOverTime, 2018.

2017 [LO-2] J. Garcia Flores, I. V. Meza Ruiz et L. A. Pineda Cortes. Golfred : Robot Experience Stories Generator (code). https://github.com/jorgegarciaflo/Golfred, 2017.

[LO-1] I. V. Meza Ruiz, G. Fuentes-Pineda, J. Garcia Flores, M. Chabouni et Z. Khezane. minhashing : Topic mining data processing and visualization (mostly on wikipedia), 2017.

Thèses et Habilitations (TH)

Thèses

2023 [TH-12] G. Felhi. Interpretable Sentence Representation with Variational Autoencoders and Attention. Thèse, Université Sorbonne Paris Nord, 2023.

2022 [TH-11] X. Zhang. Multitask Deep Learning for Joint Syntactic and Semantic Structure Prediction. Thèse, Université Sorbonne Paris Nord, 2022.

2021 [TH-10] J. Lopez Espejel. Résumé automatique abstractif des textes médicaux longs avec un Transformer à multi-encodeurs et évaluation de résumé du domaine général avec wikiSERA. Thèse, Université Sorbonne Paris Nord, 2021.

[TH-9] S. Zevio. Découverte et enrichissement de connaissances à partir de textes pour la recherche d'experts. Thèse, Université Sorbonne Paris Nord, 2021.

2019 [TH-8] I. Garrido Marquez. Dynamics in semantic annotation, a perspective of information access system. Thèse, Université Sorbonne Paris Nord, 2019.

[TH-7] S. D. Hernandez-Perez. Détection de topiques et leur similarité dans les textes scientifiques. Thèse, Université Sorbonne Paris Nord, 2019.

2018 [TH-6] C. Corro. Lagrangian Based Approaches for Lexicalized Tree Adjoining Grammar Parsing. Thèse, Université Sorbonne Paris Cité, 2018.

[TH-5] N. Firoozeh. Semantic-oriented Recommendation for Content Enrichment / Recommandation sémantique pour l'enrichissement de contenus textuels. Thèse, Université Sorbonne Paris Cité, 2018.

[TH-4] P. Holat. Fouille de motifs et modélisation statistique pour l'extraction de connaissances textuelles. Thèse, Université Sorbonne Paris Cité, 2018.

[TH-3] O. Seminck. Modèles cognitifs et computationnels de la résolution des pronoms. Thèse, Université Paris Cité, 2018. Thèse encadrée par Adeline Nazarenko.

2017 [TH-2] I. Bannour. Recherche d'information sémantique : graphe sémantico-documentaire et propagation d'activation. Thèse, Université Sorbonne Paris Cité, 2017.

[TH-1] E. Hassan. Reconnaissance d'expériences vécues dans les avis d'utilisateurs : une méthode basée sur les événements. Thèse, Université Sorbonne Paris Cité, 2017.

Habilitations

2018 [HR-1] E. Cartier. Dynamique lexicale des langues : éléments théoriques, méthodes automatiques, expérimentations en français contemporain. HdR, Université Sorbonne Paris Cité, 2018.

Rapports de contrats

2019 [RC-1] C. Nédellec, A. Nazarenko, F. Andre, C. Balivo, B. Daille, A. Drouot, J. Garcia Flores, L. Freyt-Caffin, N. Gandon, J. Lafait, N. Le Ba, C. Martin, N. Morcrette, S. Respingue-Perrin, S. Rennes et P. Zweigenbaum. Recommandations sur l'analyse automatique de documents : acquisition, gestion, exploration. Rapport Technique, Comité pour la science ouverte, 2019.

Publications antérieures à l'arrivée au LIPN

Les références comportent « ↑ » et correspondent à des publications antérieures à l'arrivée au LIPN, non référencées par ailleurs.

Signification des soulignements :
- *simple* : enseignant-chercheur ou chercheur titulaire,
- *double* : doctorant,
- *pointillé* : post-doc, ingénieur ou autre situation.

Articles dans des revues avec comité de lecture et chapitres de livres (ACL)

Articles dans des revues internationales avec comité de lecture sélectif

2017 [↑RI-1] A. Grezka et L. Zhu. Du figement au défigement : la reconnaissance de néologismes polylexicaux. *Etudes de Linguistique Appliquée*, 2/2017(186) :185–196, 2017.

Articles dans des revues nationales ou internationales

2018 [↑RE-3] J. Raad, W. Beek, N. Pernelle, F. Saïs et F. van Harmelen. Détection de liens d'identité erronés en utilisant la détection de communautés dans les graphes d'identité. *Ingénierie des Systèmes d Infrmation*, 23(3-4) :61–88, 2018.

[↑RE-2] J. Raad, N. Pernelle, F. Saïs, J. Dibie, L. Ibanescu et S. Dervaux. Comment représenter et découvrir des liens d'identité contextuelle dans une base de connaissances : Application à des données expérimentales en sciences du vivant. *Revue d'Intelligence Artificielle*, 32(3) :345–372, 2018.

2017 [↑RE-1] A. Grezka. Morfetik, un dictionnaire morphologique : illustration avec le lexique de la perception. *Écho des études romanes*, 13(1) :89–99, 2017.

Communications avec actes (ACT)

Communications dans des conférences internationales avec comité de lecture

2019 [↑CI-5] N. Ballier, T. Gaillat, A. Simpkin, B. Stearns, M. Bouyé et M. Zarrouk. A Supervised Learning Model for the Automatic Assessment of Language Levels Based on Learner Errors. *The 14th European Conference on Technology Enhanced Learning (EC-TEL 2019), Delft, The Netherlands, September 16-19, 2019*, volume 11722 de *Lecture Notes in Computer Science*, pages 308–320, 2019.

[↑CI-4] T. Mecharnia, L. Chibout Khelifa, N. Pernelle et F. Hamdi. An Approach Toward a Prediction of the Presence of Asbestos in Buildings Based on Incomplete Temporal Descriptions of Marketed Products. *The 10th International Conference on Knowledge Capture (K-CAP 2019), Marina Del Rey, CA, USA, November 19-21, 2019*, pages 239–242, 2019.

2018 [↑CI-3] J. Raad, W. Beek, F. van Harmelen, N. Pernelle et F. Saïs. Detecting Erroneous Identity Links on the Web Using Network Metrics. *The 17th International Semantic Web Conference (ISWC 2018), Monterey, CA, USA, October 8-12, 2018*, volume 11136 de *Lecture Notes in Computer Science*, pages 391–407, 2018.

2017 [↑CI-2] J. Raad, N. Pernelle et F. Saïs. Detection of Contextual Identity Links in a Knowledge Base. *The 2017 Knowledge Capture Conference (K-CAP 2017), Austin,TX, USA, December 4-6, 2017*, pages 8 :1–8, 2017.

[↑CI-1] D. Symeonidou, L. Galárraga, N. Pernelle, F. Saïs et F. M. Suchanek. VICKEY : Mining Conditional Keys on Knowledge Bases. *The 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 21-25, 2017*, volume 10587 de *Lecture Notes in Computer Science*, pages 661–677, 2017.

Autres articles dans des conférences ou workshops à comité de lecture

2019 [↑CO-7] M. Atencia, J. David, J. Euzenat, L. Ibanescu, N. Pernelle, F. Saïs, E. Thiéblin et C. Trojahn. Discovering Expressive Rules for Complex Ontology Matching and Data Interlinking. *The 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (OM@ISWC 2019), Auckland, New Zealand, October 26, 2019*, volume 2536 de *CEUR Workshop Proceedings*, pages 199–200, 2019.

[↑CO-6] N. Ballier, T. Gaillat, M. Zarrouk, A. Simpkin, M. Bouyé et B. Stearns. A Franco-Irish project for the automatic identification of criterial features in learners of English. *The 27th EUROCALL conference (EUROCALL 2019), Université de Louvain la Neuve, Belgium, August 28-31, 2019*, Louvain la Neuve, Belgium, 2019.

[↑CO-5] T. Gaillat, N. Ballier, M. Zarrouk, A. Simpkin, B. Stearns et M. Bouyé. Investigating Criterial Features of Learner English and Predicting CEFR Levels in French Learners of English. *The 27th EUROCALL Conference (EUROCALL 2019), Louvain-la-Neuve, Belgium, August 28-31, 2019*, 2019.

[↑CO-4] A. Khajeh, N. Pernelle et F. Saïs. Generating Referring Expressions from Knowledge Graphs. *The 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (OM@ISWC 2019), Auckland, New Zealand, October 26, 2019*, volume 2536 de *CEUR Workshop Proceedings*, pages 193–194, 2019.

[↑CO-3] T. Mecharnia, N. Pernelle, L. Chibout Khelifa et F. Hamdi. Approche de prédiction de présence d'amiante dans les bâtiments basée sur l'exploitation des descriptions temporelles incomplètes de produits commercialisés. *30èmes Journées francophones d'Ingénierie des Connaissance (IC 2019), Toulouse, France, July 2-4, 2019*, pages 144–157, 2019.

2017 [↑CO-2] D. Chistaras, N. Pernelle et F. Saïs. Approche numérique pour l'invalidation de liens d'identité (owl : SameAs). *28èmes Journées francophones d'Ingénierie des Connaissances (IC 2017), Caen, France, July 3-7, 2017*, pages 50–55, 2017.

[↑CO-1] J. Raad, N. Pernelle et F. Saïs. Détection de liens d'identité contextuels dans une base de connaissances. C. Roussey, éditeur, *28èmes Journées francophones d'Ingénierie des Connaissances (IC 2017), Caen, France, July 3-7, 2017*, pages 56–67, 2017.

Édition d'ouvrages collectifs

2018 [↑ED-2] P. Chapman, D. Endres et N. Pernelle. *The 23rd International Conference on Conceptual Structures (ICCS 2018), Edinburgh, UK, June 20-22, 2018*, volume 10872 de *Lecture Notes in Computer Science*. Springer, 2018.

2017 [↑ED-1] A. Grezka et M.-H. Viguier. *Figement en mouvement et complexité en devenir : regards sur quelques problèmes liés aux structures défigées et complexes*. Numéro 2/2017 Études de linguistique appliquée. klincksieck (didier erudition jusqu'en 2003), 2017.

3.3 Domaine 4. Inscription des activités de recherche dans la société

3.3.1 Référence 1. L'équipe se distingue par la qualité et la quantité de ses interactions avec le monde non-académique.

En raison de son expertise reconnue dans la conception de modèles d'apprentissage et de gestion de connaissances dédiés aux données textuelles et aux graphes de connaissances, l'équipe est régulièrement sollicitée pour des collaborations industrielles.

Sur la période on compte 3 prestations de service réalisées pour des entreprises : Crowd Prediction (2021–2023 et 2022-2023), API Conseil (2021-2022).

Au niveau doctoral, on dénombre 5 thèses sur des financements industriels dont 3 thèses CIFRE avec les entreprises Pixalione (2016-2019, thèse soutenue en 2018) Thalès DMS (2021-2024, thèse en cours), Iniciativas (2021-2024, thèse en cours). Une thèse CIFRE avec la BPCE et l'équipe A3 a par ailleurs démarré en avril 2023 suite à

un co-encadrement d'un stage Master de recherche avec cet établissement en 2022. En outre, une thèse s'est déroulée au CEA et a été financée à 100% par cet établissement (thèse soutenue en 2021).

La quatrième thèse industrielle (soutenue en 2021) a été réalisée dans le cadre d'un projet FUI pour un financement LIPN de 213k€ (2017-2020). Il s'agit du projet FUI PCUv2 dont le consortium comprenait, outre le LIPN, les entreprises Smile, Proxem, Armadillo et Wallix. Ce projet a aussi permis le financement d'un postdoc.

3.3.2 Référence 2. L'équipe développe des produits à destination du monde culturel, économique et social.

L'équipe développe aussi des recherches à destination du monde culturel, social ou de la santé. Un travail en collaboration avec un psychologue de l'université de Galway sur la détection des émotions chez les personnes atteintes d'autisme est en cours. L'enjeu est de permettre le développement d'applications personnalisées pour la communication et les interactions sociales de ce public spécifique. Du fait d'une expression particulière de l'émotion par cette population, les systèmes actuels ne sont pas efficaces. Centrée sur l'interaction personne-ordinateur, au lieu de l'interaction classique personne-thérapeute, un jeu de données multimodales comprenant 57012 exemples annotés qui permettra d'entraîner des modèles d'apprentissage, est en cours d'annotation et sera disponible pour les chercheurs, notamment les chercheurs dans le domaine de la santé.

Une autre action développée par l'équipe s'intéresse au développement d'une plateforme d'accès à des textes philosophiques (https://tal.lipn.univ-paris13.fr/bni/). Ce travail s'inscrit dans le champ des humanités numériques et associent plusieurs éditeurs pour fournir les fonds documentaires. La plateforme a une ambition de démocratisation de l'accès aux philosophes et à leurs oeuvres. Elle offre des fonctionnalités d'accès riches par l'exploitation de réseaux sémantiques. Deux types de réseau sont construits : des grilles de lecture par auteur qui présentent une vision sypnoptique des oeuvres et des thésaurus qui recensent les termes du domaine permettent ainsi des navigations à ces différents niveaux. D'autres fonctionnalités vont enrichir les possibilités de navigation dans les oeuvres. Actuellement, la plateforme indexe les oeuvres de Adam Smith et Francis Hutcheson qui seront complétées par de nouveaux auteurs ainsi que par des oeuvres qui traitent des auteurs (préfaces, correspondance...). Elle a été présentée à la fête de la sciences en 2017.

Chapitre 4
Trajectoire de l'équipe RCLN

Sommaire

4.1 SWOT **53**
4.2 Trajectoire **55**
4.2.1 Langues peu dotées 55
4.2.2 Algorithmes pour l'analyse syntaxique et sémantique de textes . . 56
4.2.3 Exploitation conjointe des modèles de langue et des graphes de connaissances 57
4.2.4 Fouille de règles pour l'enrichissement des graphes de connaissances 58
4.2.5 Stratégie partenariale locale, nationale et internationale, et Stratégie sociétale 59

Nous présentons tout d'abord une analyse SWOT afin d'évaluer nos forces, faiblesses et opportunités, puis présentons nos objectifs pour la prochaine période.

4.1 SWOT

- **Forces et Atouts**

  — *Complémentarité des compétences* : les compétences des membres de l'équipe

sont particulièrement complémentaires à la fois en terme d'objectifs puisqu'elles vont de l'analyse syntaxique à la construction et l'enrichissement de graphes de connaissances et en terme de méthodes utilisées (linguistique de corpus, apprentissage profond, apprentissage symbolique et optimisation combinatoire). Les recrutements sur la période (1 fin 2019 et 1 en 2020) ont permis de consolider l'axe Acquisition de connaissances, et cela permet d'envisager encore plus solidement des perspectives de recherche articulant traitement automatique des langues et graphes de connaissances.

— *Visibilité* : le positionnement original de l'équipe sur les problèmes difficiles en langue, sur les langues sous-dotées ou langues de spécialités ainsi que ces compétences sur la construction et l'enrichissement de graphes de connaissances lui confère une bonne visibilité nationale et internationale qui est attestée par sa participation active dans des instances nationales, des comités de programme de conférences internationales de haut niveau et des organisations de workshops internationaux.

— *Présence de nombreux jeunes chercheurs* : malgré le faible nombre d'HDR, et l'absence de Master de l'USPN portant sur les thématiques de RCLN, le nombre de doctorants (co-)encadrés est plus élevé que sur la période précédente (17 à l'ED Galilée, et 26 co-encadrements au total sur la période), avec des recrutements d'étudiants au niveau local, national et international. Une attractivité certainement due en partie à la participation de l'équipe au labex EFL.

— *Projets et contrats de recherche* : le positionnement de l'équipe à l'interface de l'informatique et de la linguistique ont permis de développer de nombreuses actions au sein du labex EFL, et de porter ou de participer à un nombre conséquent de projets académiques et industriels (3 ANR, 1 ITN H2020, 3 projets bilatéraux, 1 FUI, 3 thèses CIFRE, contrats d'expertise auprès d'industriels).

— *Valorisation logicielle et mise à disposition de ressources* : la présence d'un IR CNRS et la participation active de deux des membres RCLN à la Dev team du LIPN permet de faciliter la réalisation et la mise à disposition de la communauté scientifique de nombreux logiciels et prototypes de recherche ainsi que de nombreuses ressources (Graphes de connaissances, benchmarks, ressource morphologique).

— *Parité* : Sur 15 membres permanents, 7 membres sont des femmes (47%), et sur la période parmi les 23 doctorants et post-doctorants recrutés, 10 sont des femmes (43%).

Trajectoire de l'équipe RCLN

- **Faiblesses**

— *Faible nombre d'HDR* : seuls les 3 PR de l'équipe sont habilités, et l'un d'entre eux a pris d'importantes fonctions au sein du CNRS depuis le début de la période. Cependant, deux habilitations de maîtres de conférences sont en cours de rédaction et leur soutenance est prévue fin 2023.

— *Absence de chercheur* : depuis 2009, l'équipe ne compte plus de chercheur permanent, ce qui limite sa capacité d'action dans un contexte où les charges et responsabilités au niveau local et national sont lourdes.

- **Opportunités**

— *Collaborations au sein du LIPN* : la présence au LIPN d'une équipe de recherche spécialisée en Apprentissage (A3) et en Optimisation Combinatoire (A0C) permet à l'équipe de mener des recherches en collaborations avec des membres de ces deux équipes. Ces collaborations ont été renforcées durant la période par trois co-encadrement de thèses dont l'une a démarré début 2023, 9 co-publications ainsi que des participations et nouveaux dépôts de projets communs (1 FUI, 1 ANR en cours, 2 ANR déposées en 2023). La constitution d'un axe transverse Science des

Données au LIPN va sans doute nous amener à pérenniser ces collaborations et explorer de nouvelles thématiques de recherche communes.
— *Projets internationaux* : les collaborations internationales actuelles de l'équipe devraient à terme déboucher sur des projets internationaux plus nombreux.
— *Nouvelles technologies* : l'apparition de nouvelles technologies et de nouvelles ressources telles que les grands modèles de langues (e.g. BERT, GPT) font émerger de nouveaux usages des outils de TAL et de nouvelles thématiques de recherche pour lesquelles les compétences complémentaires de l'équipe en TAL et en graphes de connaissance peuvent être mises à profit.

- **Menaces**

— *Labex EFL* : il existe une incertitude quant au renouvellement du Labex après 2024 et cela pourrait entraîner une diminution importante des ressources de l'équipe. Cependant RCLN participe très activement à la proposition de renouvellement du Labex qui sera déposée fin 2023.
— *Recrutement et encadrement des doctorants* : l'absence de master adossé aux thématiques de l'équipe complique le recrutement des doctorants et le faible nombre d'HDR incite les MCF de l'équipe à mener des co-encadrements en collaboration avec d'autres laboratoires.

4.2 Trajectoire

introduction

Depuis la création de l'équipe et jusqu'aux années 2000, l'équipe RCLN a essentiellement développé des méthodes de TAL basées sur des modèles symboliques, et des recherches en ingénierie des connaissances. A partir des années 2010, les approches développées ont évolué vers l'apprentissage automatique que ce soit pour l'analyse ou l'acquisition automatique de connaissances. Les progrès réalisés dans le domaine de l'apprentissage profond au cours des quinze dernières années ont entraîné une amélioration indéniable des performances dans le traitement du langage naturel (NLP) ou dans celui des graphes de connaissance, qu'il s'agisse de tâches telles que la traduction automatique, l'analyse syntaxique, la génération automatique de textes ou la complétion de graphes. Depuis la dernière période, l'équipe RCLN s'est en grande majorité tournée vers des recherches basées sur ces méthodes. Dans le précédent rapport HCERES, les perspectives mentionnaient (1) le développement d'approches pour l'analyse syntaxique et sémantique, et en particulier l'utilisation des réseaux de neurones, et des méthodes issues de l'optimisation combinatoire, (2) l'annotation sémantique et exploration textuelle et (3) l'acquisition de connaissances à partir de textes. Nous avons pu développer des approches innovantes pour chacune de ces thématiques et avons appliqué certaines d'entre elles au thème transverse concernant la littérature scientifique. En revanche, le départ du professeur responsable du deuxième axe transverse proposé concernant la lecture automatique nous a amené à abandonner cet axe de recherche.

Nos objectifs de recherche actuels concernent essentiellement le traitement automatique de langues sous-dotées, les algorithmes pour l'analyse syntaxique et sémantique de textes, l'exploitation conjointe de textes et de graphes de connaissances et la fouille de règles pour l'enrichissement de graphes de connaissances.

4.2.1 Langues peu dotées

Nous souhaitons poursuivre nos recherches sur la définition d'approches de TAL dédiées aux langues peu dotées afin de mettre à disposition des ressources numériques pour ces langues et parce que ces langues posent souvent des questions intéressantes

sur le plan linguistique et sur leur traitement automatique. Le traitement de la variation linguistique représente un défi important à la fois en linguistique et en TAL, en particulier pour les langues sous-dotées, et les modèles développés dans le cadre d'une langue et de ses variations peuvent servir de référence pour l'annotation d'autres langues peu documentées. Ainsi, malgré une tradition écrite multi-séculaire, l'arménien manque cruellement de ressources numériques. La langue arménienne comprend l'arménien classique (5e-10e siècle après J.-C.), l'arménien moyen (11e-17e siècle) et l'arménien moderne (17e siècle jusqu'à aujourd'hui) avec ses deux normes : l'arménien moderne oriental et l'arménien moderne occidental. Outre les deux normes, le continuum linguistique arménien comprend divers dialectes, ainsi que des formes vernaculaires mais toutes les variétés utilisent un unique alphabet. Nous sommes partenaires du projet ANR DALHI (Digitizing Armenian Linguistic Heritage : Armenian Multivariational Corpus and DataProcessing) qui vient de débuter et qui vise à construire pour la première fois une plateforme numérique linguistique libre d'accès, pour l'ensemble des variations de la langue arménienne. Six corpus annotés concernant chacun différentes variétés de cette langue sont en cours de constitution par des membres de l'INALCO et de spécialistes de l'arménien et l'équipe RCLN se focalisera sur la définition de méthodes pour le traitement des langues avec des données rares, afin de proposer des alternatives aux tendances axées sur les données quantitatives. Les développements récents dans le domaine de l'apprentissage profond en TAL ont montré l'intérêt des modèles linguistiques pré-entraînés, basés sur les plongements de mots contextuels, tels que BERT et m-BERT. Ces modèles sont entraînés sur une grande quantité de textes non annotés, en tenant compte du contexte bidirectionnel. Les modèles de langue multilingues tels que m-Bert présentent deux avantages : ils évitent de former un modèle pour chaque variété et tirent parti des similitudes structurelles pour apprendre des caractéristiques interlinguistiques, ce qui peut être considéré comme une forme de transfert des langues riches vers les langues plus pauvres en ressources. Les modèles qui ont été développés pour l'arabe standard serviront de base pour le transfert vers des dialectes moins dotés et nous étudierons en particulier le caractère indispensable d'une modélisation explicite de la morphologie en la comparant à une modélisation implicite par les modèles de langues (multilingue) utilisés pour le transfert. Nous appliquerons ces approches à la langue arménienne pour étudier le caractère généralisable de nos conclusions mais aussi pour développer, efficacement, des modèles pour l'arménien pour laquelle nous collectons actuellement les données dans le cadre du projet DALiH.

Trajectoire de l'équipe RCLN

4.2.2 Algorithmes pour l'analyse syntaxique et sémantique de textes

Les approches basées sur de l'apprentissage profond sont à la fois gourmandes en données, sont difficilement appliquées aux langues à faibles ressources, et sont gourmandes en calculs puisqu'elles nécessitent un nombre sans cesse croissant de paramètres pour améliorer les modèles précédents. En outre, les manipulations symboliques telles que la logique et les structures, qui sont de nature combinatoire, sont encore mal prises en compte par la nature continue des réseaux neuronaux. À titre d'exemple, même les modèles récents d'analyse syntaxique des constituants ou des dépendances nécessitent encore un algorithme combinatoire pour construire un arbre syntaxique valide, puisque la solution du classifieur seul peut générer des graphes non valides. Enfin, ces modèles sont monolithiques et manquent d'un mécanisme d'accord entre les tâches en plus du partage des paramètres. Par exemple, l'analyse syntaxique avec des grammaires lexicalisées (divisées en choix de la structure de parangon correcte pour chaque mot dans le contexte, et en construction d'une analyse syntaxique), l'extraction de relations (identification d'entités et de prédicats, et établissement de liens entre eux) peuvent naturellement être exprimées comme des sous-tâches qui doivent s'accorder. Le plus souvent, ces sous-tâches pourraient être simples à résoudre de manière

isolée, mais la recherche d'un accord ajoute de la complexité qui rend les tâches globales difficiles ou irréalisables. Nous proposons de poursuivre notre collaboration avec l'équipe AOC pour étudier comment faire prédire à un réseau neuronal une solution approximative d'un processus d'optimisation. Par exemple, les algorithmes d'optimisation itératifs couramment utilisés dans le cadre lagrangien peuvent être remplacés par des algorithmes amortis et semi-amortis, d'une manière similaire à l'amortissement des distributions de propositions dans les méthodes variationnelles probabilistes [1] et plus généralement dans l'optimisation continue [**?**].

4.2.3 Exploitation conjointe des modèles de langue et des graphes de connaissances

Les grands modèles de langues tels que la famille des GPT-x et les technologies associées offrent des possibilités sans précédent à la société dans son ensemble, et plus particulièrement aux chercheurs : elles peuvent permettre de communiquer des résultats scientifiques de manière plus efficace, via une interaction homme-machine, ce qui peut libérer du temps pour se concentrer par exemple sur le développement d'idées novatrices. Toutefois, en dépit de l'engouement qu'elles suscitent, ces technologies ont de nombreuses limites : ces modèles ne peuvent pas être librement adaptés par les individus en fonction de leurs besoins, les textes générés peuvent contenir des informations incorrectes ou biaisées, elles s'appuient sur des modèles de langues de très grande taille, qui sont difficiles à constituer, et enfin, il existe des préoccupations éthiques quant à leur utilisation abusive pour générer de la fausse science et du plagiat.

Au cours des dernières années, nous avons assisté à l'émergence de graphes de connaissances (KG) qui décrivent explicitement les connaissances de la recherche dans le but de permettre à des systèmes intelligents de soutenir et accélérer les recherches scientifique. Ces ressources caractérisent des entités telles que les tâches, les méthodes, les techniques d'évaluation, ou des objets d'études telles que des protéines ou des produits chimiques), leurs relations, et les acteurs impliqués ou documents pertinents (e.g. chercheurs, organisations, articles). Leur utilisation peuvent permettre d'aider à construire des systèmes intelligents pour expliquer, conforter ou réfuter des assertions générées automatiquement ou pour fournir une aide à la rédaction intelligente de texte. Certains de ces graphes sont alimentés par crowd-sourcing (par exemple, ORKG, UMLS, Nanopublications), tandis que d'autres sont générés automatiquement à partir du texte et des métadonnées des articles de recherche tels que AI-KG que nous avons co-développé, CSO, ou encore TKG. Comme beaucoup d'autres KG, ceux qui décrivent les entités liées à la recherche souffrent d'incomplétude. Aussi nous poursuivrons nos recherches sur les problématiques de complétion automatique de graphes. Le problème de l'incomplétude des graphes de connaissances est généralement traité par des techniques de prédiction de liens ou de classification de triplets, qui ont donné de bons résultats dans plusieurs domaines. Ces méthodes utilisent souvent des modèles de plongement de graphes (e.g. TransE, RotatE, ComplEx), des caractéristiques basées sur les chemins, ou des réseaux de neurones graphiques (GNN). Nous étendrons les travaux de recherche actuels sur les GNN, en examinant des modèles plus avancés tels que les Graph Attention Networks (GAT) [5] et leurs variantes. Par exemple, la particularité de prédire les relations dans un graphe tel que AI-KG est que les connexions de ce graphe ont tendance à se concentrer autour de certains noeuds importants (graphes en étoile). Par conséquent, il existe un certain biais en faveur des noeuds ayant un degré plus élevé qu'il faut pouvoir contrôler. Pour limiter les risques d'utilisation abusive des modèles génératifs dans la production de contenu académique, le contenu généré doit être expliqué. L'explicabilité est un moyen de valider le contenu généré par des recherches dans des sources de connaissances. Dans le cas de textes scientifiques, cette méthode peut être mise en oeuvre en reliant le texte à d'autres travaux, avec de multiples avantages : l'explication de la suggestion elle-même, une suggestion de citation, et également un aspect didactique permettant d'apprendre qui a travaillé sur des sujets

similaires. Nous souhaitons mettre en oeuvre des modèles de langage qui combinent des informations lexicales et sémantiques, en tirant parti de graphes de connaissances linguistiques bidirectionnels profonds, et en utilisant conjointement les graphes de citations et les graphes de connaissances scientifiques.

Face à la prolifération des contenus générés automatiquement et aux risques d'inconduite scientifique que cela représente pour le monde académique, il est essentiel de développer des méthodes de détection efficaces pour disqualifier les textes et en particulier les articles de recherche générés automatiquement. Nous avons récemment proposé un benchmark qui s'inscrit dans ce cadre et qui inclut des textes générés automatiquement mais partiellement manipulés. Nous avons considéré la détection comme une tâche de classification binaire et testé divers algorithmes de classification. Les résultats montrent que les modèles de classification récents existants ont pu fournir une précision maximale de 70,2 % sur notre ensemble de données. Ce résultat montre que ce problème reste ouvert. Aussi, nous envisageons de développer des méthodes originales orientées vers la détection de textes et de fragments générés automatiquement, en tirant éventuellement parti des connaissances pour détecter les contradictions et les phrases hors contexte.

4.2.4 Fouille de règles pour l'enrichissement des graphes de connaissances

Découvrir des règles de la logique du premier ordre dans de grands graphes de connaissances (KG) a été abordé dans différents travaux qui diffèrent selon les biais de langage utilisés, ou les stratégies d'exploration de l'espace des règles [2, 4, 3]. Ces règles peuvent permettre d'expliciter des motifs implicites dans les graphes, de détecter des données erronées ou de faire de la complétion de graphes. Bien que les approches existantes soient efficaces, elles ne sont pas en mesure de découvrir des règles qui impliquent des contraintes définies sur des valeurs numériques ou temporelles. Or, nous pensons que ces contraintes peuvent être très pertinentes dans des domaines tels que les sciences sociales, la santé publique ou la littérature scientifique en général. Nous nous intéresserons à la découverte de règles impliquant des contraintes numériques telles que des intervalles dans de grands graphes de connaissance. Nous chercherons à sélectionner efficacement les prédicats numériques et les contraintes numériques en utilisant des techniques supervisées qui distinguent le mieux les prédictions correctes et non correctes d'une règle détectée par un système de fouille de règles efficace existant.

Trajectoire de l'équipe RCLN

De plus, nos recherches sur la fouille de règles dans les graphes de connaissances se sont récemment focalisées sur l'extraction de règles causales. Ces approches ont montré que l'effet causal peut être évalué pour des graphes de connaissances par une estimation des contre-factuels en se basant sur des individus similaires. Cependant, nous avons montré que les méthodes de calcul des similarités de deux descriptions impactent fortement le nombre et la qualité des règles causales découvertes (i.e. une similarité trop contrainte réduit fortement le nombre de règles mais garantit une meilleure qualité de la règle) [CI-29]. Nous étudierons comment des approches de fouille de règles causales peuvent se baser sur des profils générés par méthodes énumératives non supervisées définies pour les graphes attribués. Nous étudierons en particulier comment ces méthodes peuvent être utilisées pour évaluer l'influence des messages alimentaires de différentes natures sur le comportement alimentaires d'élèves (collaboration qui vient de débuter avec le LEPS de l'USPN et l'équipe A3), et donc dans un cadre où l'on dispose à la fois des descriptions des élèves, des relations qu'ils entretiennent et des descriptions des messages alimentaires qu'ils reçoivent.

4.2.5 Stratégie partenariale locale, nationale et internationale, et Stratégie sociétale

Nous souhaitons poursuivre et renforcer nos partenariats avec l'équipe AOC et A3 au sein du laboratoire. Pour cela, nous favorisons le dépôt de projets communs et le co-encadrement de thèses, et participerons activement à l'axe transverse Sciences des données afin de faire émerger de nouvelles thématiques de recherche communes.

La convention LabEx avec l'IDEX prendra fin en décembre 2024. Les opérations financées par le LabEx et portées par des membres de l'équipe RCLN ont clairement contribué à soutenir les thématiques de recherche de l'équipe (recrutements de doctorants, de post-doctorants, financement de bourses de mobilité, de matériels, nouvelles collaborations nationales et internationales). Aussi, l'ensemble des partenaires est pleinement investi pour répondre à l'appel à soumission de nouveaux projets LabEx de l'IDEX Paris Cité. Sur le plan thématique, le projet de l'axe Analyse sémantique computationnelle (axe 5) dans lequel nous sommes impliqués consiste à tirer parti du nouveau paradigme marqué par les évolutions technologiques récentes que sont les modèles de langue dont la dernière incarnation est chatGPT. L'axe souhaite étudier comment les modèles de langue (en général) peuvent contribuer à apporter de nouvelles connaissances, à étudier ces nouveaux modèles pour mieux les comprendre et à développer des activités applicatives tant en direction de la société en général que pour la communauté en sciences du langage, activités telles que l'extraction d'information, la vérification de l'information générée, l'étude et la conception de systèmes d'apprentissage pour les langues rares et peu dotées. Aussi, les perspectives de l'équipe en NLP s'inscrivent pleinement dans ce nouveau projet.

En janvier 2023, l'USPN vient de déposer un projet d'alliance européenne. L'objectif du projet UNINOVIS - Data for L.I.F.E. est de créer une alliance de sept établissements d'enseignement supérieur publics répartis dans toute l'Europe (University Sorbonne - Paris Nord, Technical University of Applied Sciences - Würzburg-Schweinfurt - Germany, University of Campania Luigi Vanvitelli - Caserta - Italy, University of Applied Sciences - Tampere - Finland, University of Applied Sciences - Kaunas - Lithuania, University of Malaga - Spain, University of Tirana - Albania). Cette alliance souhaite combiner son expertise afin d'améliorer les capacités et les compétences des européens pour comprendre, analyser, gérer, utiliser et protéger les données, et ainsi contribuer à certains des défis majeurs de l'Europe du 21e siècle. UNINOVIS vise à devenir l'université européenne leader dans l'apprentissage des sciences des données dans tous les domaines, de la médecine aux biotechnologies, de l'ingénierie à la linguistique, des affaires aux sciences sociales, de l'économie aux sciences humaines. L'équipe RCLN s'est impliquée dans le montage du projet et s'investira dans les actions pédagogiques menées en collaboration avec les autres universités afin que ses compétences en IA et en NLP puissent être mises à profit afin que les étudiants de tout domaine puissent mieux comprendre et manipuler les données issues des nouveaux outils de NLP ou des graphes de données. De même, la problématique des nouveaux usages des modèles de langue pour la littérature scientifique, nous a conduit à être partenaire d'un projet Horizon CL4 2023 qui est en cours d'évaluation.

Nous formons une équipe de nature interdisciplinaire puisqu'elle inclut des spécialistes en linguistique de corpus et des informaticiens. Nous souhaitons non seulement travailler sur des applications qui bénéficient à des experts issus d'autres domaines que l'informatique, comme nous le faisons en partie (linguistique, santé publique, droit, environnement), mais aussi renforcer les aspects interdisciplinaires de nos recherches en collaborant plus intensivement avec des experts de domaine, comme c'est le cas pour la collaboration initiée avec LEPS sur des thématiques de science de l'éducation et de nutrition.

Références complémentaires

[1] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[2] J. Lajus, L. Galárraga, and F. M. Suchanek. Fast and exact rule mining with AMIE 3. In A. Harth, S. Kirrane, A.-C. N. Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, and M. Cochez, editors, *The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings*, volume 12123 of *Lecture Notes in Computer Science*, pages 36–52. Springer, 2020.

[3] C. Meilicke, M. W. Chekol, M. Fink, and H. Stuckenschmidt. Reinforced anytime bottom up rule learning for knowledge graph completion. *CoRR*, abs/2004.04412, 2020.

[4] S. Ortona, V. V. Meduri, and P. Papotti. Rudik : Rule discovery in knowledge bases. *Proc. VLDB Endow.*, 11(12) :1946 ?1949, aug 2018.

[5] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Partie III
Acronymes

A³ Apprentissage Artificiel et Applications

ANR Agence Nationale de la Recherche

AOC Algorithmes et Optimisation Combinatoire

ASPC Alliance Sorbonne-Paris-Cité

CALIN Combinatoire, Algorithmique et Interactions

CIFRE Conventions Industrielles de Formation par la Recherche

CNRS Centre National de la Recherche Scientifique

Complexités Axe transverse Complexités

CR Chargé de Recherche

DR Directeur de Recherche

DIM Domaine d'Intérêt Majeur

EA Equipe d'Accueil

EFL Labex *Empirical Foundations of Linguistics*

FR Fédération de Recherche

FUI Fonds Unique Interministériel

IG Institut Galilée

IR Ingénieur de Recherche (CNRS)

IUTV IUT de Villetaneuse

L2TI Laboratoire de traitement et transport de l'information, USPN - EA 3043

LabCom Laboratoire commun (Appel PIA 4)

LabEx Laboratoire d'Excellence (appel Grand Emprunt)

LAGA Laboratoire Analyse, Géométrie et Applications, USPN - UMR 7539

LIPN Laboratoire d'Informatique de Paris-Nord, Université Sorbonne Paris Nord - UMR 7030

LLSHS Lettres, Langues, Sciences Humaines et des Sociétés

LoVe Logique et Vérification

MathSTIC Fédération de Recherche MathSTIC, Université Sorbonne Paris Nord - FR 3734

MCF Maître de Conférences

Prag Professeur agrégé

PU Professeur des Universités

RCLN Représentation des Connaissances et Langage Naturel

RFSI Réseau Francilien en Sciences Informatiques

UFR Unité de Formation et de Recherche

UMR Unité Mixte de Recherche

USPN Université Sorbonne Paris Nord