

Rapport 2007-2012

Projet 2014-2018

**Représentation
des Connaissances
et Langage Naturel**

Section	Sub-sections	Page
Partie I	Le LIPN	page 5
Partie II	The LIPN at a glance	7
	Présentation générale	13
	Activités de recherche	19
	Production scientifique, valorisation, dissémination	35
	Rayonnement	43
	Formation à la recherche	49
	Publications	55
Publications antérieures à l'arrivée au LIPN	75	
Partie III	Acronymes	page 83

Partie I

Le LIPN

page 5

The LIPN at a glance

Scientific activities
Environment
Doctoral studies

7

Partie II

Equipe RCLN

page 11

Présentation générale

Overview
Membres de l'équipe

13

Activités de recherche

Bilan 2007-2012
Analyse AFOM/SWOT
Perspectives et projet

19

Production scientifique, valorisation, dissémination

Publications
Contrats de recherche
Valorisation de la recherche
Dissémination de la recherche

35

Rayonnement

Comités d'organisation et d'édition
Expertises
Invitations

43

Formation à la recherche

Masters, Ecoles doctorale et post-doctorale
Thèses et Habilitations
Stages de master, post-doctorats

49

Publications

55

Publications antérieures à l'arrivée au LIPN

75

Partie III

Acronymes

page 83

The image features a teal background. A white circle is positioned on the left side, partially overlapping the teal area. A vertical black line is located to the right of the circle, extending from the top to the bottom of the circle's vertical span. The text 'Partie I' and 'Le LIPN' is centered within the white circle.

Partie I
Le LIPN

Chapitre 1

The LIPN at a glance

The LIPN¹ (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 Paris 13 which is part of the PRES Sorbonne Paris Cité.

The laboratory comprises 5 teams with 78 permanent staff (researchers or professors and associate professors), 8 technical and administrative staff, as well as PhD students and postdocs. It has overall more than 140 members. The LIPN has experienced an impressive growth during the last few years, which allowed for strengthening its major 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 (presidency of the CNRS section, CNU — national universities council, 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, and conduct collaborative research with colleagues worldwide.

1. <http://lipn.univ-paris13.fr>

1.1 Scientific activities

The laboratory is structured into 5 teams, 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 non-linear 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 having skills in a variety of aspects of combinatorics (analytic, algebraic and bijective combinatorics), interested in complexity of algorithms, trying to determine their behaviour in average or in distribution, in fine analysis of data structures, and also studying problems or using methods issued from physics. The team is organised in two axes: one focuses on the analysis of algorithms and combinatorial structures, the other is devoted to combinatorial physics.

- **LCR (Logique, Calcul et Raisonnement): Logic, Computation and Reasoning**

The LCR team tackles different aspects of computational and software models, from theory to applications. The specification and verification group addresses methodologies for the development of specifications, and verification through modular, compositional, distributed and parameterised approaches. The logic, theory of computation and programming languages group has expertise in computational issues of logical systems and formal semantics of programming languages: use of tools and methodologies issued from Linear Logic to analyse resource sensitive properties of programs, study graph formalisms for representing proofs ; algebraic and geometric foundations of computation.

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

The RCLN team is interested in natural language for its expressive capacity, and in knowledge representation, in particular for its connection to semantic analysis and text understanding. It tackles corpus analysis and semantic annotation of texts. Search and structuration of semantic web knowledge is achieved by the design of ontologies and knowledge patterns. Text understanding and knowledge acquisition are integrated in a virtuous circle between language, data and knowledge.

1.2 Environment

The LIPN has many research cooperations with its environment: other laboratories at University Paris 13, but also in the PRES Sorbonne Paris Cité, with industry, etc.

- **MathSTIC research federation, and University Paris 13 laboratories**

Research collaborations with other laboratories of University Paris 13 have been favoured over the years, in particular with LAGA² (UMR 7539, mathematics laboratory), L2TI³ (signal and image processing).

This led us to recently set up a new research federation in Mathematics and Information Technology (MathSTIC⁴), gathering members of LAGA, LIPN, and L2TI laboratories, in order to enhance cross-fertilisation in three of their major areas of expertise:

- Optimisation and learning applied to digital contents;
- High-performance computing, distributed systems;
- Mathematical physics, statistical physics, combinatorics.

- **Regional setting: PRES Sorbonne Paris Cité, and EFL LabEx**

The PRES Sorbonne Paris Cité comprises 8 universities and colleges. It includes University Paris 5, and University Paris 7, which both have computer science laboratories (LIPADE, LIAFA, and PPS). The PRES offers support for research projects between these laboratories.

The team RCLN at LIPN is a key actor in the EFL LabEx (Empirical Foundations of Linguistics). Its multidisciplinary 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. Approximately 15 foreign researchers are invited for a month at our laboratory each year to carry collaborative research. Cooperation projects also involve laboratories in Germany, Italy, Norway, Poland, Tunisia, Canada, Brazil, Chile, Uruguay, Singapore, Vietnam.

- **Industrial collaborations**

Most research areas of the LIPN also participate in collaborations with the industry, mainly through research projects (ANR, FP7, FUI, FEDER) and CIFRE PhDs. Some large national or European collaborative projects involve both academic and industrial research laboratories. 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 3 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 “Galilée” doctoral school, which addresses four major scientific areas:

1. mathematics, computer science, signal processing;
2. physics, materials, engineering sciences;

2. <http://www.math.univ-paris13.fr/laga/>

3. <http://www-l2ti.univ-paris13.fr/>

4. <http://mathstic.univ-paris13.fr/>

1.3 Doctoral studies

3. health sciences, medicine, human biology, chemistry;
4. ethology.

LIPN is part of the first group.

The doctoral school organises mandatory courses, and allocates grants to doctoral students. In 2012, six of them were allocated to the LIPN (out of 25 for the whole doctoral school).

• **PhDs and Habilitations**

Over the 2008–2012 period, 65 PhD defences took place. Half of the doctoral students were financed by a grant from the doctoral school, and the others through research projects, industry, or international support.

During the same period, 15 habilitations (Habilitations à diriger les Recherches) were delivered.

The laboratory organises an oral presentation halfway through the PhD (after 18 months), in front of a jury comprising an external member, who often is to become one of the reviewers. The aim is thus to assess the work that has already been done, and confirm the future directions of the doctoral work. Doctoral students benefit from a broader view of their work as well as advice from the jury.

• **Masters courses**

A masters programme in computer science is delivered at University Paris 13. It currently features two specialities:

- EID² (Exploration Informatique des Données et Décisionnel): Data Mining, Analytics, and Knowledge Discovery;
- 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, ADAMA 2012, Logics and Interactions 2008 and 2012, Proof Theory 2012, Alea 2012, AAFD, JPOC/ISCO were co-organised by LIPN members. Some of these thematic schools take place in University Paris 13, while others are elsewhere in France or abroad.



Partie II
Equipe RCLN

Chapitre 1

Présentation générale

1.1 Overview

Knowledge is very generally transmitted through natural language, which justifies that 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, texts understanding involves knowledge and reasoning. The RCLN team is interested in Natural Language for its expressive capacity, and in Knowledge Representation, in particular for its connection to semantic analysis and text understanding.

Evolution of the team composition The team was strongly renewed during the last period with, at first in 2007 and 2008, two arrivals who filled the previous deficit, and in 2009, three departures related to promotions and changes of laboratory, which led the team to work short staffed between 2009 and 2011. The investment of the remaining members has nevertheless allowed to carry out major projects during that period: Quaero, FP7 ONTORULE and the preparation of the LabEx EFL proposal. The four recruitments in 2011 and 2012 and the two expected in 2013 help boosting the team dynamics despite one retirement in 2012 and two scheduled for 2013.

1.1 Overview

Results The work carried out by the RCLN team is at the same time fundamental and applied, with significant results on the following topics:

- *Fine-grained understanding of text*: model of norm-based understanding, computational model of argumentation, methodology for the formalization of regulatory texts;
- *Corpus semantics*: robust approach of automatic annotation, experiments in semantic annotation, formal analysis of the complexity of manual annotation tasks;
- *Text-based knowledge acquisition*: platform for the acquisition of ontologies and semantic resources from texts, text-to-knowledge formalization methodology, method for subcategorization frames acquisition, protocols and campaigns for evaluating acquisition technologies;
- *Access to textual content*: discourse-based approach for restructuring documents, tools for multi-document summarizing, algorithms uncovering classes of behavior in folksonomies, semantic search engine.

The common objective was to develop methods to model, to various degrees, the contents of texts and documents, to facilitate browsing and searching information or to build domain models. More fundamental research on "fine-grained understanding" feeds the topic "Corpus semantics", even if the necessary robustness of corpus processing imposes simplifications. "Knowledge acquisition" relies on the results of the first two topics to work out ontologies based on textual knowledge. All these technologies contribute to improving access to textual content.

Project Taking advantage of the previous results, of the maturity of semantic web technologies, of the growth of the linked open data cloud and of the recent recruitments, the RCLN team plans to reorient its research along the following axes for the next period:

- *Corpus semantic analysis and annotation* with a focus on the integration of semantic calculus, its representation in the form of a consistent set of annotations anchored in the source text and the linguistic/extra-linguistic semantics interface.
- *Knowledge discovery and structuration for the Semantic Web* As the RCLN ontology building methodology is now mature, knowledge acquisition will be considered in a new perspective, that of the semantic web, which raises new challenges in knowledge discovery, collaboration, semantic heterogeneity and integration with the linked data. The RCLN team will benefit from the expertise of a new Professor in the semantic web field.
- *Semantic information retrieval* This theme will also be strengthened by the participation of new RCLN members. The objective is to progress towards an "open world" approach taking into account various sources of knowledge (texts, multiple ontologies, linked data).
- *Syntax-discourse integration* The recent recruitments strengthen the RCLN skills in computational linguistics, which allows to consider the different levels of linguistic analysis (syntax, semantics, discourse) and their integration, especially in the processing of specialized languages.

An additional transversal project links these different areas: the goal is to create a virtuous circle between language, knowledge and data, in which text analysis relies on data and knowledge but also contributes to their acquisition.

Context and visibility The research RCLN team fits in the scientific project of the LabEx « Empirical Foundations of Linguistics » (EFL) in which the RCLN team is involved, especially in that of the Strand 5 « Computational semantics ». It will also rely on the A³ and AOC teams expertise when introducing machine learning or combinatorial optimization technics in text mining.

At the national level, the RCLN has an unusual combination of Natural Language Processing and Knowledge Engineering skills. RCLN is well recognized in these core fields (e.g. organization of TALN, TIA and IC conferences) and begins to be in semantic information retrieval.

At the European and international levels where the association between language processing and knowledge engineering is more traditional, the originality of the RCLN approach lies in the place left to humans in knowledge modeling and in text interpretation. The goal is to gain automation at all levels, using formal and probabilistic approaches, while identifying the limits of what can be automated and designing tools for what cannot be. The thematic renewal, the increased hybridization of methods, and the strengthening of the team with new members arrival should increase its international visibility.

Other activities The RCLN research includes a significant dimension of experimental and implementation work. In most cases, the proposed methods and tools are tested and evaluated on real use cases.





Besides scientific publications, the dissemination of RCLN research results towards the socio-economic and cultural world was made through large collaborative projects and Cifre conventions, but also through the involvement in Cap Digital, the French business cluster for digital content and services.

A special effort has been made in terms of research training (tutorials in major international conferences, participation and coordination of master programs, PhD supervision) but the next challenge is to exploit the diversity of expertises of the LabEx EFL to build a new high-level pluridisciplinary master for students to be prepared to face the challenges of empirical semantics and knowledge representation raised by the semantic web and linked data.

1.2 Membres de l'équipe

1.2.1 Membres permanents

(membres ou, en italique, anciens membres, au 01/10/2012)

	<i>Nom</i>	<i>Prénom</i>	<i>Situation</i>	<i>Institution</i>
	Audibert	Laurent	MCF	IUTV
	Buscaldi	Davide	MCF	IUTV
	<i>Després</i>	<i>Sylvie</i>	MCF	IG
	Gayral	Françoise	MCF	IUTV
	Gangemi	Aldo	PU	IG
	<i>Hamon</i>	<i>Thierry</i>	MCF	IG
	Kayser	Daniel	PU	IG
	Le Roux	Joseph	MCF	IG
	Lévy	François	PU émérite	IUTV
	Nazarenko	Adeline	PU	IG

1.2 Membres de l'équipe

Présentation générale



Poibeau	Thierry	CR	CNRS
Recanati	Catherine	MCF	IG
Rozenknop	Antoine	MCF	IG
Salotti	Sylvie	MCF	IUTB
Szulman	Sylvie	MCF	IUTV
Zargayouna	Haifa	MCF	IUTV

1.2.2 Doctorants au 1er oct. 2012



<i>Nom</i>	<i>Prénom</i>	<i>Situation</i>	<i>Financement</i>
Bannour	Ines	Contrat doctoral	Université Paris 13
Bannour	Sondes	CDD sur contrat	Oséo Programme Quaero
Ben Abbès	Sarra	CDD sur contrat	Oséo Programme Quaero
Fort	Karën	CDD ingénieur & ATER	INIST & Université de Lorraine
Guissé	Abdoulaye	AM	Université Paris 13
Mimouni	Nada	Contrat doctoral	FUI Projet Légilocal
Omrane	Nouha	CDD sur contrat	FP7 Projet ONTORULE
van Puymbrouck	Jonathan	AM	Université Paris 13

1.2.3 Autres personnels temporaires du 1/1/2007 au 1/10/2012

<i>Nom</i>	<i>Prénom</i>	<i>Situation</i>	<i>Début</i>	<i>Fin</i>	<i>Financement</i>
Athier	Mickael	Ingénieur d'études	1/2010	8/2010	ANR projet Dafoe4App
Bentebibel	Rim	Post-doctorante	1/2009	11/2009	ANR projet Dafoe4App
Bossard	Aurélien	Doctorant	10/2006	7/2010	Université Paris 13
Bouffier	Amanda	Doctorante	10/2004	10/2008	Université Paris 13
Bourgeois	Laure	ATER	10/2006	8/2007	Université Paris 13
Boutouhami	Sara	Doctorante	11/2006	11/2010	Projet MICRAC
Cailliau	Frederik	Doctorant	3/2005	12/2010	CIFRE Sinequa
Ceausu-Dragos	Valentina	ATER	2/2007	8/2007	Université Paris 13
Derivière	Julien	Ingénieur d'études	6/2004	2/2007	FP6 projet Alvis & FUI projet Infom@gic
Généreux	Michel	Post-doctorant	12/2007	7/2009	FUI projet Infom@gic
Hamon	Olivier	Doctorant	3/2006	12/2010	CIFRE ELDA
Jacques	Marie-Paule	Post-doctorante	10/2006	08/2007	ANR projet TextCoop
Kamenhoff	Nicolas	Doctorant	10/2007	10/2011	Epitech

1.2 Membres de l'équipe

Nouioua	Farid	Doctorant	10/2003	4/2007	Université Paris 13
Ma	Yue	Post-doctorante	7/2008	2/2012	Oséo Programme Quaero
Messiant	Cédric	Doctorant	10/2006	11/2010	DGA
Mondary	Thibault	Doctorant	10/2006	05/2011	Université Paris 13
		& Ingénieur de recherche	6/2011	12/2013	& Oséo Programme Quaero
Moreau	Erwan	Post-doctorant	5/2008	12/2010	ANR projet Crotal & Oséo Programme Quaero
Presutti	Valentina	Chercheur invité	9/2012	12/2012	CNRS
Roux-Thomas	Aurélien	Ingénieur d'études	5/2010	9/2010	FP7 projet ONTORULE
Weissenbacher	Davy	Doctorant	10/2003	9/2007	Université Paris 13
Zargayouna	Haïfa	Post-doctorante	1/2007	8/2007	FUI projet Infom@gic

**Présentation
générale**

1.2.4 Evolution de l'équipe

L'équipe RCLN s'est fortement renouvelée sur la période avec d'abord, en 2007 et 2008, des arrivées qui sont venues combler le déficit antérieur, puis, en 2009, des départs liés à des promotions et changement de laboratoire, ce qui a conduit l'équipe à fonctionner en effectif réduit entre 2009 et 2011. Les quatre recrutements de 2011 et 2012 permettent de relancer la dynamique de l'équipe et d'envisager sereinement les départs en retraite de 2012 (F. Lévy) et programmés pour 2013 (S. Szulman et D. Kayser), d'autant que deux nouveaux recrutements sont prévus en 2013.

• Permanents

Départs

- S. Després, MCF, promotion PR, 2009
- T. Hamon, MCF, changement de laboratoire, 2009
- T. Poibeau, CR, changement de laboratoire et promotion DR, 2009
- F. Lévy, PU, départ en retraite et éméritat, 2012

Arrivées

- S. Després, MCF, mutation (MCF Univ. Paris 5), 2007
- H. Zargayouna, MCF, postdoc LRI/Inria, 2008
- J. Le Roux, MCF, postdoc LIFM, 2011
- D. Buscaldi, MCF, postdoc IRIT, 2012
- A. Gangemi, PU, Senior researcher CNR (IT), 2012
- Ingénieur de recherche CNRS, prévu en déc. 2012

Chapitre 2

Activités de recherche

2.1 Bilan 2007-2012

L'équipe RCLN s'intéresse à la langue pour son pouvoir expressif, et à la représentation des connaissances, notamment en tant qu'outil mis au service du traitement du langage naturel. Ses travaux, qui comportent des recherches à caractère fondamental et des recherches plus appliquées, se sont organisés autour de quatre thèmes :

- la compréhension fine de textes,
- la sémantique de corpus,
- l'acquisition de connaissances à partir de textes,
- l'accès à l'information textuelle,

C'est la combinaison de compétences en traitement automatique des langues et en ingénierie des connaissances qui confère à l'équipe RCLN un positionnement original au plan national. L'équipe RCLN s'inscrit par ailleurs dans le projet scientifique du LabEx EFL auquel elle participe, notamment dans celui de l'axe « Analyse sémantique computationnelle » que coordonne A. Nazarenko. Elle bénéficie également de l'expertise de l'équipe A³ pour introduire des techniques d'apprentissage dans le traitement de données textuelles.

2.1 Bilan 2007-2012

Ces recherches comportant une part importante d'expérimentation, elles se sont accompagnées d'un effort significatif de développement, ainsi que de test et d'évaluation dans le cadre de projets collaboratifs (voir sections 3.3.2 et 3.2).

L'équipe RCLN a pris une part active dans l'animation des communautés nationales du traitement automatique des langues et de l'ingénierie des connaissances par l'organisation de conférences, la participation à des comités de rédaction et de programme, à travers divers jurys et expertises ainsi que par l'animation du LabEx « Fondements empiriques de la linguistique » (EFL, voir section 4.1).

Le rayonnement des travaux et résultats de l'équipe a été assuré par des publications scientifiques mais aussi, auprès du monde socio-économique, à travers d'importants projets collaboratifs, par des thèses Cifre, et par l'implication dans le pôle de compétitivité Cap Digital (voir sections 3.2, 3.3.1 et 4.2.1). L'équipe RCLN a également poursuivi son effort de formation à la recherche à travers des tutoriels dans des conférences internationales, la coordination de spécialités de master, la conception de nombreux modules de master, l'accueil et le suivi de stagiaires de recherche et de doctorants (voir section 5.1).

Il convient de répondre aux remarques émises par l'AERES en 2007, qui portaient sur l'articulation avec la linguistique, la visibilité et les publications, l'encadrement doctoral et l'évaluation des technologies développées. L'*interface avec la linguistique* a été nettement renforcée dans le cadre du LabEx EFL. L'équipe RCLN a poursuivi ses travaux sur corpus mais en intégrant des modèles sémantiques formels¹ et elle projette d'exploiter dans la prochaine période les résultats de la linguistique formelle, laquelle progresse de son côté en couverture et adaptabilité. En termes de *visibilité*, l'équipe RCLN est désormais bien reconnue au plan national. Au niveau international, l'équipe, du fait de son effectif réduit, a focalisé ses efforts sur la collaboration avec des partenaires européens dans le cadre de projets, notamment le projet FP7 ONTORULE. La réorientation qualitative de l'activité de *publication* doit être poursuivie, beaucoup de travaux expérimentaux étant encore publiés dans des workshops spécifiques. L'*activité d'encadrement doctoral* a été soutenue, en dépit du départ de membres habilités. Un réel effort a été fait pour l'*évaluation* des méthodes et outils proposés, par l'organisation d'évaluations [RI-9, CI-62] et la participation à des campagnes d'évaluation internationales (voir par ex. [CO-7, CI-30, CO-91]).

2.1.1 Compréhension fine de textes

L'analyse de la sémantique fine des textes vise à modéliser l'interprétation qu'un humain fait des textes et notamment ce qu'il en infère. Il s'agit d'une recherche de long terme de l'équipe RCLN. Trois questions ont été plus spécifiquement étudiées.

- **Raisonnement sur les causes et les normes**

Parmi les mécanismes d'interprétation mis en jeu dans la compréhension d'un texte, l'équipe RCLN a mis l'accent sur les processus d'identification de relations causales entre événements, d'explications causales et d'attribution de responsabilités. *Résultats* :

- Analyse de la diversité des raisonnements causaux (objectifs, notions fondamentales utilisées, propriétés) [CI-31];
- Modèle du raisonnement causal dans la compréhension de textes [RI-10], qui a été testé sur un corpus de constats d'accidents.

- **Modéliser le raisonnement argumentatif**

Le texte ne sert pas uniquement à informer, il est aussi construit pour présenter les faits sous un jour favorable aux objectifs de l'énonciateur. L'équipe RCLN a cherché

1. Il s'agit dans ce cas d'une sémantique extra-linguistique, modélisée par exemple sous la forme d'ontologies de domaine.

à intégrer le calcul argumentatif dans les processus de traitement automatique des langues, en mettant l'accent sur l'organisation et la hiérarchisation des arguments.

Résultats :

- Modèle procédural permettant de générer des textes argumentés selon différents points de vue (mécanismes d'insinuation, de justification, de référence à une norme coutumière, etc.) [RE-4];
- Validation de ce modèle sur un ensemble de textes réels dont l'impact argumentatif a été augmenté [RI-19].

• **Interpréter les textes réglementaires**

Un autre aspect de la compréhension fine a été abordé dans le projet FP7 ONTORULE dans le but de formaliser des textes réglementaires en un ensemble de règles métiers puis de vérifier la conformité de procédures ou de produits à la réglementation en vigueur. Au-delà de l'enjeu applicatif, il s'agit d'apprécier l'apport des techniques de traitement automatique des langues à l'interprétation sémantique des textes en situation réelle (complexité des textes, criticité des connaissances extraites). *Résultats :*

- Méthodologie d'aide à l'interprétation des textes réglementaires et à l'acquisition de règles métiers [CI-64];
- Plateforme Semex implémentant cette méthode pour guider le travail d'extraction, de normalisation et de modélisation des connaissances [CI-51, CI-55];
- Application de cette méthodologie et de la plateforme associée sur différents cas d'usage industriels dans le cadre du projet ONTORULE ².

**Activités de
recherche**

2.1.2 Sémantiques de corpus

Les recherches menées autour de l'analyse sémantique de corpus se sont progressivement concentrées sur des tâches d'annotation sémantique. Sans rompre avec les travaux d'analyse de corpus menés antérieurement par l'équipe et dans l'état de l'art (par ex. [2]), cette approche met l'accent sur l'ancrage dans le texte de la représentation sémantique produite, quelles que soient la richesse et la complexité de cette dernière : au lieu de se substituer au texte, l'annotation se surajoute à lui. Elle formalise aussi le lien entre les niveaux de représentation linguistique et extra-linguistique, les annotations faisant référence à un modèle sémantique formel extérieur (par ex. une ontologie de domaine). Cette conception de l'analyse comme annotation a émergé de travaux portant à la fois sur l'annotation en tant que telle – manuelle ou automatique – et sur l'intégration de différents niveaux d'annotation dans un même processus d'interprétation sémantique.

• **Annotation manuelle de corpus**

Dans le contexte du programme Quaero qui suscite un effort de création de corpus annotés sans précédent en France, l'équipe RCLN a initié une réflexion méthodologique sur la conduite de ces campagnes d'annotation qui étaient gérées jusque-là de manière dispersée et pragmatique. Ce travail fait l'objet d'une thèse en collaboration avec l'INIST qui est en charge de certaines campagnes d'annotation de Quaero. *Résultats :*

- Cadre d'analyse et métriques permettent de mesurer *a priori* la complexité d'une tâche d'annotation selon 6 dimensions [CI-65*];
- Méthodologie de gestion des campagnes d'annotation manuelle couvrant l'analyse des tâches, le contenu des guides d'annotation [CO-53], l'évaluation de la qualité d'annotation produite [CO-82], le choix des outils d'aide à l'annotation [CO-66] et le pilotage des annotateurs [CI-59].

2. Voir <http://ontorule-project.eu/news/news/ontorule-showcase.html>

• Annotation automatique de corpus

L'annotation automatique de corpus repose généralement sur l'application de différents modules de traitement sur un même texte. Il en résulte une superposition de couches d'annotation qui sont par nature hétérogènes, multiples, parfois concurrentes, inévitablement bruitées, mais sur lesquelles l'analyse impose de raisonner. Les travaux de l'équipe RCLN ont visé à corriger les erreurs ou les incohérences tout en exploitant au mieux la diversité des annotations produites par les outils. *Résultats* :

- Modèles bayésien [CI-2] et probabiliste [RI-23] qui permettent d'exploiter la diversité des annotations, aussi incertaines soient-elles, produites par des modules de traitement automatique des langues et qui ont été appliqués avec succès à la résolution de l'anaphore pronominale dans des textes de génomique [TH-3];
- Formalisation du processus d'annotation sémantique à l'aide de règles d'inférence de différents types [CI-50] dont la cohérence est mesurable;
- Méthode d'apprentissage statistique permettant d'entraîner un annotateur sémantique à partir d'un petit corpus annoté (travail en cours de publication);
- Plate-forme d'annotation UIMA³ [LO-2] qui devient progressivement une base logicielle centrale pour les tâches d'annotation au sein de l'équipe RCLN.

Ces résultats ont d'autant plus d'importance que l'intégration de modules de traitements hétérogènes devient un enjeu majeur pour l'analyse sémantique. C'est l'un des défis auxquels le LabEx EFL a choisi de s'attaquer.

• Annotation sémantique étendue

L'annotation sémantique vise généralement à enrichir les textes d'un ensemble d'annotations qui font référence à un modèle sémantique et s'analysent au regard de ce dernier. Le plus souvent, dans l'état de l'art, ce modèle est une ontologie et les segments de textes annotés s'interprètent comme des mentions d'instances de concepts [4, 1] mais l'équipe RCLN a montré l'intérêt d'une annotation plus riche, notamment sur les cas d'usage du projet ONTORULE. *Résultats* :

- Modèle d'annotation prenant en compte tout type d'entité ontologique (instances de concepts mais aussi concepts, rôles et instances de rôles) [RE-5];
- Modèle d'annotation augmenté au regard d'une base de connaissance articulant une ontologie et un ensemble de règles métiers [CI-55];
- Modèle de navigation dans un texte et le(s) modèle(s) sémantique(s) associé(s) : les expériences ont montré que ces liens d'annotation construisent « des modèles sémantiques documentés » qui servent à la traçabilité, au contrôle de cohérence et à la mise à jour des connaissances extraites [CI-64].

2.1.3 Acquisition de connaissances à partir de textes

L'acquisition de connaissances et la construction de ressources sémantiques (ontologies, thesaurus, etc.) à partir de textes est un thème ancien de l'équipe RCLN, à la frontière du traitement automatique des langues, de l'analyse sémantique de corpus et de l'ingénierie des connaissances. Les résultats obtenus sur la dernière période concernent à la fois la méthodologie d'acquisition dans son ensemble et certaines de ses étapes. Un effort a été mené en parallèle pour évaluer l'apport des méthodes d'acquisition.

• Méthodologie d'acquisition

La méthodologie Terminae de construction d'ontologies et de ressources sémantiques à partir de textes a été développée au sein de l'équipe RCLN à la fin des années

3. <http://uima.apache.org/>

1990 sur la base des travaux du groupe TIA⁴, mais elle a été récemment réexaminée, explicitée et enrichie. *Résultats* :

- Spécification de la plateforme OpenSource Dafoe de construction d'ontologies de grande taille à partir de textes [CI-33] et mise en évidence des trois structures de connaissances en jeu dans le processus d'acquisition ;
- Révision de la méthode Terminae qui repose désormais explicitement sur l'élaboration progressive de connaissances de plus en plus structurées et contraintes ;
- Nouvelle version de l'outil Terminae⁵, qui est désormais une application eclipse RCP pouvant intégrer des composants externes comme l'éditeur d'ontologies "neon toolkit"⁶ et être elle-même intégrée à d'autres outils.

• **Extraction des indices textuels**

L'acquisition de connaissances à partir de textes s'appuie en premier lieu sur l'exploitation du matériau textuel. L'équipe RCLN a pris en considération différents types d'indices, pour développer de nouvelles méthodes d'extraction et/ou mesurer leur apport à différentes tâches d'acquisition. *Résultats* :

- Méthode de filtrage des termes par les entités nommées : les expériences menées dans le cadre d'ONTORULE ont montré que les entités nommées jouent le rôle de marqueurs de domaine [CI-53] et facilitent le filtrage des autres éléments textuels ;
- Algorithme d'extraction des relations sémantiques [CO-69] : à l'opposé des approches à base de patrons, cette méthode fait émerger des relations sémantiques candidates à partir d'une analyse distributionnelle en corpus.
- Système d'acquisition de cadres de sous-catégorisation [RI-14] et lexique de sous-catégorisation pour le français [CI-23] : l'évaluation de cette ressource et sa comparaison à d'autres montrent l'apport des méthodes automatiques dans la constitution de lexiques, l'adaptabilité de tels systèmes et leur aptitude à acquérir des ressources « spécialisées » rapidement.

**Activités de
recherche**

• **Normalisation d'un réseau terminologique**

La normalisation d'un réseau terminologique consiste à filtrer les éléments non pertinents, regrouper ceux qui ont des sens proches et distinguer les différents sens des unités ambiguës. Il est essentiel d'outiller ce processus complexe et souvent fastidieux qui suppose de fréquentes consultations des textes sources. *Résultat* :

- Méthode d'aide à la normalisation terminologique reposant sur la notion de réseau terminologique pondéré et un ensemble d'opérations de normalisation (thèse en cours de rédaction).

• **Conceptualisation d'une ontologie**

La conceptualisation vise à créer et structurer un ensemble de concepts, et l'on s'appuie souvent sur le regroupement des unités textuelles en classes sémantiques pour ce faire. C'est une étape critique dans le processus de construction d'une ontologie, que l'équipe RCLN a étudiée sous différents angles. *Résultats* :

- Mise en évidence du rôle des entités nommées dans le travail de conceptualisation [CI-53], alors que l'état de l'art ne prend en compte que leur rôle dans la création d'instances ;
- Mesure de similarité distributionnelle adaptée aux unités composées que sont souvent les termes [CO-84] : les expériences menées ont montré que ces mesures servent au filtrage de termes sans pour autant permettre de construire de véritables classes de termes ;

4. Terminologie et Intelligence Artificielle

5. <http://lipn.fr/terminae/>

6. <http://neon-toolkit.org/>

2.1 Bilan 2007-2012

- Adaptation de l’analyse formelle de concepts à la classification de termes [TH-11], avec des paramétrages pour contrôler la taille, la complexité et l’interprétabilité du treillis généré.

En dépit des résultats obtenus, les deux dernières approches s’avèrent difficiles à mettre en œuvre pour aider le travail de conceptualisation. Les entités nommées, en revanche, sont désormais prises en compte, et à tous les niveaux, dans la méthode Terminae.

• **Evaluation des méthodes d’acquisition**

Des recherches ont été menées pour définir des protocoles et outils adaptés à l’évaluation des méthodes et outils d’acquisition. *Résultats* :

- Architecture et méthodologie d’évaluation pour le traitement automatique des langues [TH-7];
- Protocole d’évaluation pour les outils d’extraction de termes : décomposition en fonctionnalités élémentaires, métriques et outil permettant d’automatiser l’évaluation [RI-9, LO-3];
- Campagnes d’évaluation organisées dans le cadre du programme Quaero [CI-62];
- Protocole d’évaluation pour d’autres tâches d’acquisition de connaissances [CI-45], notamment les méthodes de construction de classes sémantiques de mots.

2.1.4 Accès à l’information textuelle

Les méthodes d’accès à l’information textuelle dépendent de l’analyse préalable du contenu textuel (segmentation en mots, typage des entités nommées, annotation conceptuelle, etc.) mais elles se caractérisent aussi par le modèle documentaire sur lequel elles reposent et les mécanismes d’exploration qu’il autorise.

• **Modéliser les documents et les collections documentaires**

Le modèle documentaire traditionnel, qui associe au texte un ensemble de descripteurs (mots, mots-clés, concepts, etc.) donnés comme représentatifs du contenu textuel, offre des possibilités d’interrogation limitées. L’équipe RCLN a exploré différents modèles alternatifs. *Résultats* :

- Modèles documentaires intégrant les structures logique [CI-6] et discursive du document [CI-43] : ces modèles ont été testés sur les guides de bonne pratique médicale pour aider les médecins dans la consultation des textes ;
- Modèles de collections documentaires prenant en compte les liens entre des documents qui, redondants et/ou complémentaires, ne peuvent pas s’interpréter isolément : [CI-44] s’appuie sur la détection des entités nommées pour regrouper des dépêches de presse en classes sémantiques ; [CO-88] représente une collection juridique comme un réseau sémantique où les documents sont liés par différents types de relations (amendement, dérivation, transposition, jurisprudence, etc.).
- Algorithme de classification folksonomique [CO-85] visant à regrouper les documents étiquetés de manière similaire par des groupes d’utilisateurs ayant eux-mêmes des profils proches.

• **Donner un accès sémantique au contenu textuel**

Les modèles documentaires trouvent leur pertinence par les méthodes d’accès au contenu qui les exploitent. Différentes approches ont été proposées au-delà de la recherche d’information traditionnelle qui retourne un sous-ensemble de la collection documentaire en réponse aux requêtes des utilisateurs. *Résultats* :

- Interface de navigation dans les grands documents [CI-6] reposant sur un double système d’indexation structurelle et sémantique (interface testée sur des guides de bonne pratique médicale) ;

- Méthode de résumé automatique multi-documents qui procède par élimination de la redondance [TH-10, CI-44] et qui a été évaluée sur la tâche « résumé d'opinions issues de blogs » de la campagne d'évaluation TAC 2008 [CI-30] et la tâche « résumé incrémental de dépêches » des campagnes TAC 2008 et TAC 2009 (classement dans le premier quart des participants); cette méthode a été intégrée dans un système applicatif permettant de visualiser les axes d'un corpus et d'extraire automatiquement les informations importantes;
- Premier modèle de recherche d'information combinant analyse formelle et relationnelle de concepts pour indexer les documents à la fois sur leur contenu sémantique et sur leurs propriétés intertextuelles [CO-88] (ce modèle est destiné à l'exploration des documentations complexes comme celle du droit);
- Terrier SIR, plateforme de test pour la recherche d'information sémantique [CO-99] développée sur la base de la plateforme de recherche d'information Terrier [3].

2.1.5 Résultats complémentaires

- Analyse syntaxique à partir de petits corpus d'apprentissage exploitant des collocations apprises en mode non-supervisé [CI-63], reposant sur l'hybridation de modèles génératifs et discriminants et permettant l'adaptation d'analyseurs syntaxiques à de nouveaux domaines;
- Méthodologie de conception de grammaires profondes lexicalisées [RI-26*].
- Etude systématique de la sémantique verbale du verbe *quitter*, modélisée par une vingtaine de schémas inférentiels [CO-79] (expériences de validation psychologique des processus d'adaptation proposés prévus dans le cadre du LabEx EFL);
- Etude interdisciplinaire sur les relations entre langage, raisonnement et connaissances [CL-8] proposant une vision de l'inférence rendant compte d'une rationalité pragmatique [RI-17], à l'œuvre notamment dans le jugement de catégorisation;
- Architecture linguistique et environnement pour la gestion des ressources linguistiques (lexiques, grammaires, corpus) [TH-8];
- Démonstration de propriétés formelles des logiques de description paraconsistantes, *i.e.* tolérant les inconsistances, en matière d'inférences [RI-25*] et de requêtes [RI-22]: ces propriétés permettent de raisonner sur des annotations inconsistantes;
- Mesures d'inconsistance dans les logiques de description paraconsistantes [RI-15], qui donnent la possibilité de chercher des diagnostics préférés.

2.2 Analyse AFOM/SWOT

• Atouts

- *Cohérence thématique* : les recherches de l'équipe RCLN articulent étroitement le traitement automatique des langues et la représentation des connaissances, ce qui constitue une originalité dans le paysage français. Les derniers recrutements renforcent cette orientation.
- *Présence de nombreux jeunes chercheurs* : le nombre de doctorants (8), d'ingénieurs et de postdoctorants (2 en moyenne) a créé une dynamique (mise en place d'ateliers, d'expériences croisées) qui contribue à la formation à la recherche et au débat scientifique.
- *Pluridisciplinarité* : avec son expertise à l'interface de l'informatique, de la linguistique et des sciences cognitives et son expérience de la collaboration avec d'autres disciplines, l'équipe RCLN jouit d'une bonne crédibilité sur des projets pluridisciplinaires comme le LabEx EFL.
- *Collaborations industrielles* : l'éventail des collaborations de RCLN offre une bonne diversité de champs d'application.

● **Faiblesses**

- *Fluctuation d’effectifs* : le départ de MCF ou CR habilités ou expérimentés (3 en 2009), les difficultés de recrutement (poste de PU laissé vacant en 2009 et 2010), les départs à la retraite prévus (3 en 2012-13) et l’implication dans des projets importants (Quaero, ONTORULE, LabEx EFL) ont fortement limité les capacités de l’équipe à investir de nouvelles thématiques et à améliorer sa communication.
- *Déficit de visibilité internationale* : les compétences de l’équipe RCLN sont insuffisamment connues au plan international, au-delà de ses partenaires de projets.
- *Absence de chercheur* : depuis 2009, l’équipe ne compte plus de chercheur permanent (CR, DR), ce qui limite sa capacité d’action dans un contexte où les charges et responsabilités liées à l’enseignement sont lourdes.

● **Opportunités**

- *Recrutements d’enseignants-chercheurs* : les recrutements effectués en 2011 et 2012 et prévus pour 2013 augmentent significativement les forces et les marges d’action de l’équipe RCLN.
- *Poste d’ingénieur de recherche CNRS* : l’arrivée d’un IR informaticien au sein de l’équipe RCLN fin 2012 vient conforter le volet expérimental de ses recherches.
- *LabEx EFL* : la création de ce LabEx assure à l’équipe RCLN un ancrage fort dans le champ de la linguistique en complément de son ancrage naturel en informatique au sein du LIPN. C’est une opportunité en termes de recherche (renouvellement des questions de TAL, renforcement des collaborations académiques) comme pour l’enseignement (projet de création d’un master recherche commun).
- *Maturité des enjeux d’analyse des contenus spécialisés* : les progrès effectués en analyse de corpus, la mise à disposition d’outils de TAL et l’importance croissante des enjeux de traitement de l’information textuelle créent un contexte très favorable pour la diffusion des méthodes d’analyse sémantique et d’ingénierie des connaissances auxquelles l’équipe RCLN travaille.
- *Pôle Math-STIC à Paris 13* : la création de ce pôle devrait permettre de renforcer les collaborations de RCLN avec les équipes A³ et AOC, qui sont précieuses dans la mesure où le TAL requiert la maîtrise de techniques d’apprentissage et d’optimisation combinatoire.

● **Menaces**

- *Master* : la suppression de la spécialité « Ingénierie des textes et contenus numériques » complique le recrutement de nouveaux doctorants, impose de leur apporter des compléments de formation et nuit au rayonnement de l’équipe. Cela rend d’autant plus critique le projet de création d’un master recherche en informatique appuyé sur le LabEx EFL.

Références complémentaires

- [1] F. Amardeilh, P. Laublet, and J.-L. Minel. Document annotation and ontology population from linguistic extractions. In *Proc. of the 3rd Int Conf. on Knowledge capture*, pages 161–168, New York, NY, USA, 2005. ACM.
- [2] S. Nirenburg and V. Raskin. *Ontological semantics*. MIT Press, 2004.
- [3] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson. Terrier information retrieval platform. In D. Losada and J. Fernández-Luna, editors, *Advances in Information Retrieval*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer Berlin / Heidelberg, 2005.
- [4] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov. Kim – a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10(3-4) :375–392, 2004.

2.3 Perspectives et projet

L'équipe RCLN oriente selon 4 axes principaux :

- l'analyse et l'annotation sémantique de corpus (section 2.3.1),
- la découverte et la structuration de connaissances pour le web sémantique (section 2.3.2),
- la recherche d'information sémantique (section 2.3.3),
- l'intégration syntaxe-discours (section 2.3.4),

avec un projet transversal (section 2.3.5) qui montre que ces différents axes sont en réalité étroitement liés : l'analyse des textes fait appel à des données et connaissances tout en contribuant à leur acquisition. L'objectif est donc de créer un cercle vertueux entre langage, données et connaissances

Contexte Au plan national, l'équipe RCLN présente une association originale de compétences en traitement automatique des langues et ingénierie des connaissances. Une dimension nouvelle, plus intégrative, vise à développer des technologies sémantiques articulant acquisition et exploitation des connaissances à l'échelle du web sémantique.

Au plan européen et international où l'association entre traitement automatique des langues et ingénierie des connaissances est plus classique, les travaux de l'équipe RCLN se distinguent par la place laissée à l'humain, dans la modélisation des connaissances et dans l'interprétation des textes. L'objectif est de gagner en automatisation, à tous les niveaux, par des approches formelles et numériques, tout en cernant les limites de ce qui est automatisé/able et en concevant des outils d'aide pour ce qui l'est pas.

En lien avec le renouvellement des thématiques et une hybridation accrue des méthodes, le renforcement de l'équipe RCLN par les recrutements récents devrait permettre d'accroître sa visibilité internationale.

Implication dans le LabEx EFL Le projet de l'équipe RCLN s'inscrit en grande partie dans les axes de recherche du LabEx « Fondements Empiriques de la Linguistique »⁷. Le défi identifié pour le traitement automatique des langues (axe 5) est l'« Analyse sémantique computationnelle », en lien avec l'analyse syntaxique et dans des contextes spécialisés où les besoins d'accès au contenu sont critiques.

Dans ce contexte, l'équipe RCLN collabore avec les équipes Alpage et LATTICE. Un des objectifs est de construire un master pluridisciplinaire de niveau européen en réponse aux défis que le web et la massification des données lancent à la représentation des connaissances et à l'analyse des contenus textuels.

Application dans les SHS et le domaine juridique Plusieurs champs d'application pourront être étudiés mais en priorité celui des SHS qui a vocation à se développer dans le cadre du Campus Condorcet et qui pose des questions complexes du fait de l'importance de la rhétorique, de la plasticité des genres et de la dynamique de l'interprétation liée à la multiplicité et aux déplacements de points de vue. Le domaine juridique, notamment, est devenu un enjeu important en termes de traitement et d'accès à l'information. L'ouverture des données dans les administrations crée de nouveaux besoins en termes de gestion de contenu et fait apparaître de nouveaux paradigmes de consommation de l'information. L'arrivée d'A. Gangemi [↑RI-2] et de D. Buscaldi [↑RI-9] renforce l'expérience de l'équipe RCLN dans ce domaine [RI-3]⁸.

2.3.1 Analyse et annotation sémantique de corpus

Les questions d'analyse sémantique ont été mises en avant dans le projet scientifique de l'axe 5 du LabEx EFL, le défi étant d'intégrer dans une même analyse les résultats

7. EFL, pour *Empirical foundations of Linguistics*

8. Voir aussi les projets Ontorule et Legilocal.

2.3 Perspectives et projet

de divers modules de traitement (extraction d'entités nommées et de relations, analyse des coréférences et des modalités, calcul de classes distributionnelles, catégorisation de textes, etc.). Avec le développement du web et des technologies sémantiques, l'analyse sémantique tend à être encodée comme une annotation sémantique, ancrée dans le texte source.

L'équipe RCLN se situe dans cette optique : analyser un texte revient à apposer un ensemble cohérent d'annotations sur le texte source, la sémantique de ces annotations étant donnée par un modèle formel extra-linguistique explicite, le plus souvent une ontologie. Coordonné par A. Nazarenko, le projet vise à combiner différents types d'annotations, alors que les premières approches mettaient l'accent sur les seules entités [12] sans tenir compte des connaissances génériques, axiomatiques, modales, etc. Il s'agit aussi de prendre en compte l'interactivité et la dynamique des processus interprétatifs en s'appuyant sur les acquis du web sémantique en matière de référence à l'identité des entités et de représentation des connaissances, du web de données en matière de réutilisation et du *crowdsourcing* pour la collaboration. L'ambition est d'arriver à une compréhension globale des besoins et mécanismes d'annotation pour proposer des méthodes et outils adaptés à différents types d'usage.

Activités de recherche

Intégrer les dimensions référentielle, textuelle et compositionnelle de l'analyse sémantique L'intégration sous la forme d'annotations des résultats de différents modules de traitement sémantique nécessite de repenser l'articulation des différentes théories ou approches du sens sur lesquelles ils reposent.

À partir des nombreuses expériences montées depuis une dizaine d'années pour l'analyse sémantique (à des degrés variables) de corpus textuels, l'équipe RCLN se donne pour objectif de proposer un cadre intégrant de manière explicite et cohérente différentes visions de la sémantique : 1) une approche référentielle guidée par le repérage d'entités du monde dans le texte et par un modèle de connaissances extra-linguistique (ontologie ou autre), 2) une approche distributionnelle et textuelle prenant en compte la détermination du sens par l'ensemble du texte et par sa structure, et 3) une approche compositionnelle plus traditionnelle produisant une représentation sémantique linguistique. Les méthodes d'analyse de corpus actuelles tendent à combiner ces différentes approches de manière très pragmatique. Comprendre leur articulation devrait permettre de maîtriser les analyses mises en œuvre et à terme d'adapter le calcul à la profondeur d'analyse et au type de texte visés pour une application donnée.

Combiner analyse et interprétation en un processus d'annotation interactif Penser l'interaction entre analyse sémantique automatique et interprétation par l'analyste humain est essentiel pour la mise en œuvre de méthodes de traitement automatique des langues en contexte réel. Rares sont les processus d'analyse sémantique entièrement automatiques : l'intervention humaine est nécessaire pour annoter des données d'entraînement, concevoir le modèle sémantique guidant l'analyse, écrire des règles d'extraction ou corriger les résultats. La qualité et la robustesse des méthodes proposées dépendent d'une juste répartition entre calcul automatique et interprétation.

Des travaux préliminaires ont montré l'importance de ces processus d'interaction [CL-5, CI-64] mais l'objectif est de définir les modalités de contrôle et de supervision de l'apprentissage à partir de données qui sont au cœur des processus d'analyse sémantique. D'un point de vue applicatif, il s'agira 1) de renouveler les approches d'extraction d'information en proposant des méthodes d'extraction légères, faciles à mettre en œuvre par des analystes humains et 2) d'outiller au mieux le travail manuel des annotateurs humains par des techniques d'apprentissage dans un contexte où les campagnes d'annotation de corpus se développent [CI-65*].

Modéliser la dynamique de l'interprétation Comme l'annotation sémantique repose à la fois sur l'analyse des données textuelles et sur des connaissances extra-

linguistiques, il faut concevoir un processus dynamique articulant révision des connaissances et analyse. Jusqu'à présent, les méthodes d'annotation sémantique développées par l'équipe RCLN reposaient sur le choix ou la construction *a priori* d'un modèle sémantique qui était ensuite exploité pour l'annotation sémantique [CI-55] mais cette approche est simplificatrice : le modèle sémantique doit être révisé quand les résultats de l'annotation font apparaître des contradictions ou des « trous » de couverture. Cela suppose de définir des critères susceptibles de déclencher le processus de révision du modèle sémantique et de procéder à cette révision en garantissant la traçabilité des annotations déjà posées pour pouvoir, le cas échéant, les mettre à jour une fois le modèle révisé.

Ce processus d'interprétation dynamique est particulièrement critique à introduire pour toutes les applications SHS où la dimension diachronique et la pluralité des points de vue mettent en échec les approches sémantiques statiques.

2.3.2 Découverte et structuration des connaissances pour le web sémantique

Activités de
recherche

La découverte et la structuration des connaissances est un enjeu majeur à l'échelle du web sémantique et du web de données (*linked data*), où les questions d'intégration de données hétérogènes sont centrales. L'équipe RCLN compte étendre ses travaux de construction d'ontologies en prenant en compte les enjeux de modularisation, de réutilisation et de collaboration. Ce projet est porté principalement par A. Gangemi⁹.

Conception d'ontologies et de patrons de connaissances La conception d'ontologies peut s'appuyer sur un corps de règles et de bonnes pratiques en utilisant des « patrons de conception ontologiques » (*Ontology Design Pattern*, ODP) [5], comme le génie logiciel ou l'architecture s'appuient sur des patrons de conception. Il a été montré par exemple que la méthodologie eXtreme Design [↑CI-12], sur des cas concrets, permet de construire plus facilement des ontologies de meilleure qualité.

Les patrons de connaissances sont un type particulier d'ODP. Ce sont des structures de connaissances réutilisables dans les tâches de modélisation et l'existence de bases de patrons disponibles [↑CL-2] favorise, de fait, leur réutilisation.

L'objectif de RCLN est de poursuivre cet effort de création de patrons de connaissances, en réutilisant des structures de connaissances provenant d'autres domaines, comme les patrons ou *frames* lexicaux présents dans les ressources lexicales [8], les microformats/données du web, les modèles de données, etc. Il s'agit aussi de développer des méthodes pour reconnaître ces patrons de connaissances dans les données. Ce projet s'appuie sur l'hypothèse d'une invariance des patrons de connaissances à travers les différentes modalités de la connaissance (structure des liens sur le web de documents et le web social, textes, cadres lexicaux, web de données, ontologies, interaction visuelle) [↑RI-7, ↑CI-17].

Un projet particulier consiste à considérer le matériau textuel comme support à la réutilisation et comme guide à l'intégration de connaissances. Il s'agit de s'appuyer sur un texte caractéristique d'un domaine pour organiser les connaissances de ce domaine (par ex. construire une ontologie). L'objectif est d'exploiter les indices textuels et des heuristiques linguistiques pour sélectionner les ODP et les intégrer dans la structure de connaissances en cours de construction ou pour aligner différents ontologies et vocabulaires les uns par rapport aux autres.

Valorisation et développement de la plateforme Terminae L'équipe RCLN compte poursuivre le travail de valorisation de la plateforme Terminae¹⁰ de construction d'onto-

9. En lien avec les activités du Semantic Technology Lab du CNRS qu'A. Gangemi dirigeait précédemment et avec lequel un partenariat est à l'étude.

10. <http://lipn.fr/terminae/>

2.3 Perspectives et projet

logies à partir de textes. Il s'agit de stabiliser une ou plusieurs versions de la plateforme correspondant à des usages identifiés, d'en assurer la diffusion *via* des *plugins* dans des éditeurs d'ontologies tels que Protege, des tutoriaux et des cas d'usage.

Le projet consiste aussi à intégrer de la collaboration dans les processus d'acquisition de connaissances. Différentes modalités sont envisagées : le travail peut être distribué de manière uniforme entre des pairs mais différents points de vue peuvent aussi collaborer à une même tâche ; le nombre des contributeurs peut varier de quelques experts à une foule d'intervenants (*crowdsourcing*) éventuellement joueurs [6], etc. Ces processus collaboratifs seront d'abord étudiés dans l'état de l'art et sur des cas d'usages concrets, l'idée étant d'enrichir progressivement la plateforme Terminae de fonctionnalités permettant de modéliser les points de vue, de suivre les changements effectués, d'intégrer des modules de connaissances et, à terme, d'acquérir des patrons de changement et d'intégration.

2.3.3 Recherche d'information sémantique

Dans le cadre large des travaux visant à concevoir des modalités sémantiques d'accès au contenu, le projet de l'équipe RCLN est principalement porté par H. Zargayouna et se trouve renforcé par l'arrivée de D. Buscaldi. Il se focalise sur trois questions particulières avec comme objectif de développer une approche de type « domaine ouvert » intégrant différentes modalités d'interrogation et d'exploration et différentes sources de connaissances (textes, ontologies, web de données), à la différence des travaux antérieurs de l'équipe majoritairement centrés sur des collections documentaires closes et des modèles sémantiques stables.

Approche logico-distributionnelle de la recherche d'information L'équipe RCLN a investi dans la mise en place d'une plateforme ouverte de recherche d'information sémantique (Terrier SIR). Celle-ci repose sur le modèle vectoriel, largement éprouvé, de Salton. Elle permet de mener des expériences comparatives de méthodes de recherche d'information sémantique et de définir de nouvelles approches exploitant à la fois le contenu textuel, avec par exemple des calculs distributionnels, et les représentations formelles qui lui sont associées *via* des raisonnements (requêtes Sparql, par ex.).

Dans la mesure où les travaux autour de l'accès sémantique au contenu manquent de cadre d'évaluation unifié, l'équipe RCLN compte poursuivre ses efforts d'évaluation¹¹ dans le domaine de la recherche d'information sémantique (organisation de campagnes, création de benchmarks, mise à disposition de ressources ouvertes), *a priori* dans une approche collaborative par des méthodes de *crowdsourcing* et/ou des jeux collaboratifs.

Ces travaux à l'intersection de la recherche d'information et du web sémantique ont vocation à être valorisées dans les projets du LabEx EFL concernant l'accès au contenu textuel. Une thèse est lancée en 2012 sur le sujet et une collaboration se met en place avec Enrico Motta et son équipe du KMi (The Knowledge Media Institute) qui ont déjà entamé un premier travail dans ce sens [4].

Recherche d'information sémantique en domaine ouvert Avec la multiplication des ontologies en ligne et l'essor du web de données, la vision monolithique où une seule ontologie sert à améliorer l'accès à l'information est remise en question. Le passage à l'échelle nécessite de prendre en compte l'hétérogénéité sémantique, avec des verrous tels que la sélection et l'alignement d'ontologies. Le but est d'avoir la plus large couverture sémantique tout en garantissant des raisonnements cohérents et en raisonnant dans un « monde ouvert ». Ce contexte imposant de repenser les modèles d'accès sémantique, l'équipe RCLN compte proposer des méthodes d'interrogation et de navigation permettant de présenter les résultats en contexte, *i.e.* selon un point de vue ou en fonction de la position du document dans la collection. Ces approches seront testées

11. Voir programme Quaero

sur des cas d'usage particuliers, comme l'accès à l'information juridique locale (projet Legilocal) ou à l'analyse de brevets de pharmacologie (programme Quaero).

Recherche d'information étendue au web de données L'impact des techniques de recherche d'information dans un contexte où les données ouvertes se généralisent impose de penser l'intégration des résultats des moteurs de recherche et des données liées du web pour permettre aux utilisateurs de passer naturellement d'une modalité de recherche à l'autre. A l'exemple de [↑CO-16, ↑CI-13], l'équipe RCLN prévoit de développer des solutions réalisant ce type d'intégration sur des cas d'usage concrets. Cela soulève un double défi : pouvoir identifier à la volée des structures de connaissances pertinentes dans le web de données en lien avec les résultats de la requête de l'utilisateur et concevoir des modes d'interaction adaptés à la recherche exploratoire et intégrant des modalités variées d'accès aux données et connaissances.

Activités de
recherche

2.3.4 Intégration syntaxe-discours

L'arrivée de J. Le Roux et les collaborations dans le cadre du LabEx EFL permettent de réinvestir le champ du traitement automatique des langues. Le projet porté principalement par J. Le Roux et A. Rozenknop vise à considérer les différents niveaux d'analyse linguistique associés à la sémantique (syntaxe, discours) et leur intégration avec cette dernière, notamment pour l'analyse des langues de spécialités. L'objectif est de combiner à terme les méthodes formelles et celles de l'analyse de corpus (section 2.3.1).

Adaptation d'analyseurs syntaxiques Bien que l'analyse syntaxique soit arrivée à maturité, on observe une baisse de la qualité des analyses quand le domaine d'évaluation s'éloigne du domaine d'apprentissage. En effet, les applications se font plutôt sur des textes de spécialité (textes juridiques, médicaux, scientifiques ...), qui diffèrent significativement, en termes de lexiques et de constructions syntaxiques mais aussi en termes de qualité de rédaction, des textes journalistiques sur lesquels les analyseurs sont traditionnellement appris. L'équipe RCLN a montré ses compétences en matière d'adaptation d'analyseurs pour des textes hors-domaine (avec les trois meilleurs systèmes à la Shared Task *Parsing the Web* 2012 [CO-91]) et l'analyse de contenus non-édités de type twitter [↑CO-22].

Le projet est d'étendre ces travaux aux textes juridiques, qui présentent certaines spécificités intéressantes : la fréquence de certains phénomènes syntaxiques complexes invite à investiguer des formalismes de représentation plus riches et l'existence de corpus parallèles, comme les directives européennes, amène à considérer de nouvelles méthodes pour extraire l'information syntaxique, en alignant les structures arborescentes [2] ou les lexiques [11]. Ces travaux s'insèrent dans les travaux du LabEx EFL sur l'apprentissage non-supervisé en TAL et devraient bénéficier de la collaboration avec A³, renforcée dans le Pôle Math-STIC.

Analyse de discours L'équipe RCLN projette de développer une méthode d'analyse discursive qui articule les niveaux d'analyse syntaxique et sémantique et qui soit adaptée aux tâches d'extraction de connaissances. L'analyse discursive se décompose généralement en deux parties : délimitation des unités élémentaires du discours (UED) puis analyse des relations qui les gouvernent. Le projet est de s'appuyer sur des corpus de référence en discours (ex. RST Discourse Treebank, Penn Discourse Treebank, ou le corpus en cours de construction pour le français dans le cadre du LabEx EFL) et d'utiliser des grammaires apprenables sur corpus, pour les deux étapes. L'originalité de l'approche envisagée par rapport à [10] et [9] est d'utiliser des méthodes à base d'annotations latentes et de réordonnement. Les premières, notamment les grammaires algébriques probabilistes à annotations latentes (PCFG-LA), ont donné lieu ces dernières années

2.3 Perspectives et projet

aux meilleurs analyseurs syntaxiques pour les langues naturelles [7]. Certaines limitations, comme l'indépendance des réécritures, peuvent être compensées en ajoutant un module de réordonnement discriminant [3], qui réévalue les structures produites à l'aune de contraintes globales [↑CO-21].

Outils d'optimisation pour l'analyse multi-niveaux L'intégration des différents niveaux d'analyse est un défi important qui a été mis en avant dans l'axe5 du LabEx EFL. Le projet de l'équipe RCLN est d'utiliser des techniques issues de l'optimisation combinatoire comme la relaxation lagrangienne pour le traitement automatique des langues [8], notamment pour construire des architectures où chaque système est lancé en parallèle et renvoie une seule solution, et où un système particulier est chargé de vérifier l'accord entre les différents modules. S'il y a accord global, la solution optimale a été trouvée ; sinon, les différents modèles sont amendés et l'architecture réévaluée, jusqu'à obtention de l'accord global. Cette approche offre la simplicité des premières architectures « chaînées » traditionnelles et évite l'explosion combinatoire, tout en permettant la révision des solutions erronées. Ce projet s'appuiera sur des collaborations avec l'équipe AOC et tirera profit des compétences de certains membres de l'équipe RCLN en optimisation combinatoire [RI-24]. Les premières expériences porteront sur l'intégration de modules d'étiquetage et d'analyse syntaxique et seront ensuite élargies à l'interface syntaxe-sémantique.

Intégration logicielle D'un point de vue technique, l'effort initié autour de la création d'une chaîne de traitement UIMA sera poursuivi. Il s'agit, en capitalisant les développements effectués dans l'équipe, de se doter d'un cadre pour monter des expériences, combiner différents modules et raisonner sur les couches d'annotation qui en résultent (annotations concurrentes, manquantes, erronées, etc.). Ce travail sera coordonné par l'ingénieur de recherche qui doit être recruté fin 2012 et bénéficiera de collaborations académiques et industrielles (LINA, Alpage, IBM, Viasema, Temis).

2.3.5 Un projet transversal entre langage, données et connaissance

Dans un contexte où la sémantique est désormais au cœur du web et du système socio-technique qui s'y développe, l'équipe RCLN projette de travailler au développement et à l'intégration de technologies pour le web sémantique et de données. Les liens existants entre les différents axes précédents seront étudiés dans le cadre d'un projet transversal sur lequel l'ensemble de l'équipe RCLN prévoit de se mobiliser et qui comporte une double dimension exploratoire et applicative.

Tirer des connaissances des gisements de textes L'équipe RCLN se donne pour défi d'enrichir le web de données à partir de la version française de wikipedia. Il s'agit d'intégrer les approches formelles et statistiques de l'analyse de contenu avec les méthodes de représentation des connaissances issues du web sémantique. Ce projet prolonge l'approche proposée par [↑CI-20]¹². Il s'agit d'exploiter les résultats d'une analyse sémantique computationnelle [1] et d'utiliser la DRT (Discourse Representation Theory, [5]) comme format intermédiaire pour alimenter les bases de connaissances du web sémantique et du web de données (RDF, OWL). Cela suppose de définir des heuristiques pour reconnaître des structures de connaissances spécifiques dans les résultats de l'analyse sémantique et de lier les éléments extraits aux données existantes, les « patrons de connaissance » ainsi mis au jour à partir de textes, lexiques, données, ontologies, etc., [↑RI-7] construisant un cercle vertueux entre langage, données et connaissance.

12. Les outils associés sont FRED (<http://wit.istc.cnr.it/stlab-tools/fred>) et Tiplao (<http://wit.istc.cnr.it/stlab-tools/tipalo>).

Ce projet transversal doit permettre, sur un problème d'école mais à « suffisamment grande échelle », de mesurer l'apport et les faiblesses des traitements automatiques de la langue (section 2.3.4 et 2.3.1), de poursuivre le travail sur la modélisation des processus de formalisation des connaissances textuelles initié dans Terminae en l'étendant à des structures de connaissances plus complexes, éventuellement dans une approche collaborative (section 2.3.2), et de tester des méthodes d'intégration de données à grande échelle (section 2.3.3).

Outiller les données publiques ouvertes La sémantique du web n'est pas un domaine de recherche purement académique. Cette sémantique est de plus en plus intégrée à des applications logicielles développées en réponse aux besoins des utilisateurs. Dans le mouvement général d'ouverture des données, les formats sémantiques du web de données sont désormais recommandés dans les administrations publiques¹³ mais leur adoption et leur généralisation nécessitent des chaînes de traitement et des outils efficaces pour la production et la transformation des données comme pour leur exploration.

Dans le prolongement de [↑CI-13]¹⁴ et du projet Legilocal, et avec l'appui de l'ingénieur de recherche qui doit être recruté fin 2012, l'équipe RCLN a comme objectif de tester sur des projets applicatifs les méthodes de conception rapide ainsi que les techniques d'analyse visuelle et d'exploration que requièrent l'intégration des connaissances et des données et l'interaction des utilisateurs avec celles-ci.

Activités de
recherche

Références

- [1] J. Bos. Wide-coverage semantic analysis with boxer. In J. Bos and R. Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications, 2008.
- [2] D. Burkett and D. Klein. Two languages are better than one (for syntactic parsing). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 877–886. Association for Computational Linguistics, 2008.
- [3] M. Collins. Discriminative reranking for natural language parsing. In P. Langley, editor, *ICML*, pages 175–182. Morgan Kaufmann, 2000.
- [4] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta. Semantically enhanced information retrieval : An ontology-based approach. *Web Semantics : Science, Services and Agents on the World Wide Web*, 9(4) :434 – 452, 2011. WS special issue on Semantic Search.
- [5] H. Kamp. A theory of truth and semantic representation. In J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof, editors, *Formal Methods in the Study of Language*, volume 1, pages 277–322. Mathematisch Centrum, Amsterdam, 1981.
- [6] M. Lafourcade and A. Joubert. A New Dynamic Approach for Lexical Networks Evaluation. In *Processings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*.
- [7] S. Petrov and D. Klein. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April 2007. Association for Computational Linguistics.
- [8] A. M. Rush, D. Sontag, M. Collins, and T. Jaakkola. On dual decomposition and linear programming relaxations for natural language processing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1–11. ACL, 2010.
- [9] K. Sagae. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *IWPT*, pages 81–84. The Association for Computational Linguistics, 2009.
- [10] R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information. In *HLT-NAACL*, 2003.
- [11] O. Täckström, R. McDonald, and J. Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. 2012.
- [12] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *Journal of Web Semantics*, 4, 2006.

13. Voir par exemple <http://spcdata.digitpa.gov.it/data.html>

14. <http://bit.ly/semanticscout> et <http://aemoo.org>

Chapitre 3

Production scientifique, valorisation, dissémination

Les travaux de l'équipe RCLN ont produit des résultats diversifiés. Ils se sont naturellement traduits par des publications scientifiques, même si cette production a suivi les fluctuations en taille de l'équipe (3 départs en 2009, dont un CR CNRS). Ils ont aussi donné lieu à des réalisations informatiques qui ont été mises en oeuvre et testées sur des projets concrets.

Les résultats et la notoriété acquise sur ses thématiques de recherche ont amené l'équipe RCLN à participer à des projets structurants (Infom@gic puis Quaero et ONTORULE) où ses compétences et son expertise étaient recherchées. Ces projets ont permis de reformuler certaines questions scientifiques et de renouveler des thématiques de recherche. Le projet ONTORULE a ainsi permis de faire émerger et reconnaître une approche de l'annotation sémantique qui se distingue de celles de l'état de l'art par sa richesse sémantique.

3.2 Contrats de recherche

3.1 Publications

	2007	2008	2009	2010	2011	2012	Total
Livres	0	1	0	0	0	0	1
Chapitres de livre	4	1	1	1	1	0	8
Revue internationale	3	3	6	8	1	5	26
Autres revues	2	2	0	1	0	0	5
Conférences internationales	17	14	13	8	7	8	67
Autres communications	17	21	22	9	14	15	98
Total	43	42	42	27	23	28	205

3.2 Contrats de recherche

Production scientifique, valorisation, dissémination

Type	Resp. sc. (Coord. ●)	Nom	Financier	Période	Durée (mois)	Subvention globale (K€)	Part LIPN (K€)
FP7							
FP7	A. Nazarenko	ONTORULE	CEE	2009-11	36	5 400	360
AII							
AII	A. Nazarenko	Quaero	Oséo	2008-13	60	99 000	800
ANR							
ANR LabEx	A. Nazarenko	EFL	ANR	2011-21	120	9 000	500**
ANR Platef.	S. Szulman	Dafoe4App	ANR	2007-10	36*	820	95
ANR MDCO	A. Rozenknop	CroTAL	ANR	2008-10	36	213	80
ANR blanc	D. Kayser	MICRAC	ANR	2006-08	325	65	
FUI							
FUI-06	A. Nazarenko	Infom@gic	Region IdF Oséo	2006-09	36*	19 300	250
FUI-09	H. Zargayouna	Legilocal	Region IdF Oséo	2009-12	36*	1 700	300
Autres							
RNTL	F. Gayral	TextCoop	ANR	2006-08	36	424	128
PPF	A. Nazarenko ●	ACTmed	MENESR	2005-08	48	200	85
MSH	T. Poibeau ●	Discours	MSH Paris-Nord	2010-12	0,5	0,25	

* indique que la durée initiale du projet a été allongée; ** indique une valeur estimée.

● ONTORULE (FP7, Large-scale integrating project, 2009-2011)

Ce projet visait à faciliter l'interaction, la gestion et le contrôle des applications métier par les acteurs du domaine en séparant une couche conceptuelle modélisant les connaissances métiers (ontologie du domaine et règles) du code de l'application, ce qui a requis de développer les modèles formels, les méthodes et les outils permettant d'acquérir et de raisonner au niveau conceptuel.

L'équipe RCLN du LIPN a apporté à ce projet sa compétence en ingénierie des connaissances textuelles. Elle a montré comment les méthodes d'acquisition à partir

de textes permettent d'exploiter les textes réglementaires [CI-51]. Elle a conçu une structure de *modèle métier documenté* où les textes sont indexés au regard du modèle sémantique et où le modèle sémantique est ancré dans les textes [RE-5, CI-55]. Cela permet d'assurer la traçabilité des connaissances métier utilisées dans les systèmes d'aide à la décision, la gestion de leur cohérence et de leur nécessaire évolution. Cette approche a été testée sur plusieurs cas d'usage [CI-54].

Partenaires *Académiques* : FUB, TUW, Fundación CTIC. *Industriels* : IBM France (coordination), Ontoprise, PNA, Audi, Arcelor-Mittal.

- **Quaero (Oséo, 2008-2013)**

Sous la direction de Technicolor (ex-Thomson), l'objectif de ce programme est de développer des outils intégrés de gestion des contenus multimédias, dont un moteur de recherche, des outils de veille scientifique et technique, etc.

L'équipe RCLN participe aux projets Recherche (*Core Technology Cluster*) et *Corpus*. Elle contribue aux travaux portant sur l'acquisition de connaissances (terminologies, ontologies) [CO-84, CO-76] et l'annotation sémantique [CO-47, CI-50, CO-78]. L'équipe est également responsable de l'évaluation des tâches d'acquisition de terminologies et d'ontologies, ce qui l'a amenée à concevoir des protocoles et/ou à mettre en place des campagnes d'évaluation [RI-9, CI-45, CI-58, CI-62].

Production scientifique, valorisation, dissémination

Partenaires *Académiques* : CNRS (INIST, LIMSI, IMMI), Inria, MIG-INRA, Institut Telecom, IRCAM, IRIT, RWTH, KIT, UJF. *Industriels* : Technicolor (coordination), Exalead, France Télécom, Yacast, Jouve.

- **Infom@gic (2006-2009)**

Ce projet structurant du pôle de compétitivité Cap Digital, visait à mettre en place, sur une période de trois ans, un laboratoire industriel de sélection, de tests, d'intégration et de validation d'applications opérationnelles des meilleures technologies franciliennes dans le domaine du traitement de l'information. Ce laboratoire s'appuie sur une plateforme UIMA commune couvrant tous les grands domaines de l'analyse d'information (la recherche et l'indexation, l'extraction de connaissances et la fusion d'informations), quelles que soient les sources de données considérées (données structurées, texte, images et sons).

L'équipe RCLN du LIPN a participé à ce projet avec l'équipe A³. Elle a surtout contribué aux tâches d'extraction de données textuelles (voir par ex. [CO-60, CI-30, CO-28]).

Partenaires *Académiques* : Universités Paris 6, Paris 8, Marne La vallée, GET, Limsi-CNRS. *Industriels* : Thales (coordination), EADS, ONERA, et une vingtaine de PME, dont Bertin, Vecsys, Sinequa.

- **LegiLocal (FUI-9 Cap Digital, 2010-2013)**

Coordonné par la société Victoires Editions, le projet LegiLocal vise à simplifier l'accès des citoyens aux données administratives, juridiques, publiques locales en proposant aux collectivités un bouquet de services qui s'intègrent aisément à leur site web en s'appuyant sur les technologies du Web Sémantique et du Web Social. L'équipe RCLN du LIPN apporte à ce projet sa double compétence en annotation sémantique et en ingénierie des connaissances. Les défis consistent à prendre en compte la dimension fortement intertextuelle des documents juridiques et différents niveaux de langue (juridique et citoyenne) dans les moteurs d'accès au contenu. Les premiers résultats concernent

3.2 Contrats de recherche

la mise en place de fonctionnalités sémantiques [CO-99] et la prise en compte de l'intertextualité caractéristique de la documentation juridique [CO-88] dans le moteur de recherche d'information.

Partenaires *Académiques* : CERSA. *Industriels* (PME) : Victoires Editions (coordination), Jamespot, Mondeca, Temis, ViaSema.

- **Labex EFL, *Fondements empiriques de la linguistique/Empirical Foundations of Linguistics* (ANR 2011-2021)**

Le labex EFL vise à ouvrir de nouvelles perspectives pour la linguistique par une approche intégrative. Il regroupe des équipes françaises de recherche de premier plan en linguistique théorique et appliquée, en linguistique computationnelle et en psycholinguistique. Grâce à des collaborations avec de prestigieuses institutions multidisciplinaires (CSLI, MIT, Institut Max Planck, SOAS ...), le projet vise à créer une École de Linguistique de Paris, un nouveau site innovant et interdisciplinaire où le dialogue entre les différentes disciplines des sciences du langage sera encouragé, avec un accent particulier sur les bases empiriques et méthodes expérimentales et une expertise précieuse sur le transfert de technologie et les applications.

La direction scientifique est assurée par Jacqueline Vaissière (Université Paris 3) assistée de plusieurs directeurs adjoints, dont Adeline Nazarenko, qui a joué un rôle moteur dans le montage du projet. Adeline Nazarenko coordonne également les activités scientifiques de l'axe "Sémantique computationnelle". Le budget alloué à ce labex dans son ensemble est de 9 millions d'euros sur 10 ans.

Partenaires : LPP, LLF, Equipe PP du LPP, Llacan , ALPAGE LATTICE , CRLAO , MII , HTL , LACITO , SEDYL , Equipe mémoire et langage du LPNCog

- **CRoTAL (ANR Masse de Donnée et Connaissances, 2008–2010)**

Le projet CRoTAL (*Conditional Random Fields* pour le TAL) dont l'objectif était l'étude et le développement de nouvelles techniques pour la manipulation de grandes masses de données textuelles, s'est focalisé sur les champs aléatoires conditionnels ou *Conditional Random Fields* (CRF). Ces modèles d'apprentissage statistique font partie à la fois de la famille des modèles graphiques et de celle des modèles d'apprentissage discriminant. Ils se situent actuellement au meilleur niveau dans plusieurs domaines d'ingénierie linguistique, comme l'extraction d'information ou la structuration des données textuelles.

Les participants du projet se sont attachés à tester, affiner et optimiser différents modèles de CRF pour les utiliser efficacement sur de très grands ensembles de données textuelles ou structurées. Au LIPN, l'étude a porté sur l'utilisation des CRF dans les graphes d'arbres. L'une des applications, l'annotation d'arbres syntaxiques en rôles sémantiques [CO-43], a donné lieu à une participation au challenge CoNLL en 2010 [CI-36] et organisation d'un atelier ATALA (25 et 26 mars 2010).

Partenaires *Académiques* : LIFO, LIFL, Inria LNE, STL, Limsi-CNRS (coordination)

- **MICRAC (ANR, 2006-2008) – Modèles Informatique et Cognitifs du raisonnement Causal**

Ce projet pluridisciplinaire reposait sur une étroite collaboration entre chercheurs en intelligence artificielle et en psychologie cognitive. Il visait à la construction d'un système informatique capable de représenter des relations causales, de reconnaître ces relations, d'expliquer des événements et d'attribuer des responsabilités à des agents.

Ce projet collaboratif a permis d'expliciter une vue multiforme de la causalité suivant ses objectifs, les notions élémentaires mises à son fondement et les propriétés qu'on lui suppose [CI-31]. Ces travaux ont donné lieu à un atelier [CO-49]. L'équipe RCLN a mis en oeuvre l'approche proposée dans son travail sur les constats d'accident [TH-1]. Des protocoles appropriés ont été construits pour valider les conjectures effectuées sur la construction d'une description à visée argumentative, protocoles qui ont été repris dans [RI-19].

Partenaires : IRIT (coordination), CRIL, CLLE-LTC

- **Textcoop, Typologies et grammaires de textes afin de rendre un TEXTE plus COOPératif en gestion de l'information (RNNTL, 2006-2008)**

Le projet TextCoop avait pour but l'utilisation de typologies textuelles et de grammaires de textes dans des systèmes d'accès à l'information textuelle (moteurs de recherche, systèmes de question/réponse, extraction d'informations).

L'équipe RCLN a travaillé à l'analyse des types de textes (manuels utilisateur, recommandations médicales, documentations techniques, inventaires, articles scientifiques, discours, etc.) [Ra-1] et à la conception de méthodes permettant de les prendre en compte dans des applications de recherche d'information. On sait en effet que la notion de type de texte influe sur certaines caractéristiques textuelles, que celles-ci concernent le lexique, la morphosyntaxe ou des modes d'organisation plus globaux.

Production scientifique, valorisation, dissémination

Partenaires Académiques : IRIT. **Industriels :** Sinequa, EADS

- **ACTMed (Plan Pluriformations pour le plan quadriennal 2005-2008)**

Ce projet réunissait trois laboratoires de l'Université Paris-Nord qui ont collaboré pour développer des méthodes d'accès à l'information textuelle médicale. Il s'agissait en priorité de concevoir les outils qui permettent aux médecins de consulter les Guides de Bonne Pratique médicale produits par l'Agence Française de Sécurité Sanitaire des Produits de Santé (AFSSAPS).

Le LIPN a montré comment une analyse du discours permet de structurer sémantiquement les Guides de Bonne Pratique médicale [CI-16, CI-43] et l'intérêt de cette structure pour le développement de fonctionnalités avancées de navigation dans ce type de documents [CI-6].

Partenaires : LimBio, LLI

- **Discours (Projet MSH 2010-2012)**

Ce projet a porté sur l'analyse du discours, des aspects linguistiques aux applications documentaires. Le but de ce projet associant linguistes et informaticiens était de contribuer à la réflexion sur les outils permettant une analyse des macro-structures du discours et sur une indexation automatique de ces passages clés dans les textes. Le financement a permis d'inviter en novembre 2010 Yufan Guo, doctorant à Cambridge qui a travaillé sur la reconnaissance et l'annotation de zones sémantiquement homogènes dans des documents de biologie.

Partenaire : Lattice

3.3 Valorisation de la recherche

3.3.1 Contrats industriels

- Convention Cifre avec la société Sinequa (2005-08) pour la thèse de Frederik Caillau ; A. Nazarenko
- Convention Cifre avec la société ELDA (2006-09) pour la thèse de Olivier Hamon ; A. Nazarenko
- Convention de thèse avec l'INIST (12/2009-12) pour la thèse de Karèn Fort ; A. Nazarenko
- Contrat de recherche EDF R&D, Expertise sur l'apport des terminologies et ontologies dans le domaine de la recherche d'information, *Étude bibliographique*, livraison 24/12/2009, budget 20 Keuros ; H. Zargayouna

3.3.2 Réalisations informatiques (logiciels), Brevets, Start-up

Annotator¹ est un plugin eclipse permettant d'annoter un texte au regard d'ontologies OWL et d'un thesaurus SKOS. L'annotation consiste à marquer les unités textuelles correspondant à des éléments de l'ontologie (concept, instance). Le thesaurus contient les unités terminologiques associées aux éléments de l'ontologie. Le texte annoté en RDFa peut être visualisé dans un navigateur. L'annotateur est intégré à l'outil SemEx et à la plateforme Terminae.

Cet annotateur a été développé par Abdoulaye Guissé, François Lévy et Sylvie Szulman.

Concordancier UIMA est un concordancier générique basé sur le langage TextMarker et la plateforme UIMA. Il prend en entrée un texte annoté linguistiquement à l'aide de la plateforme UIMA (fichier .xmi) et un Type System (fichier .xml) contenant la description des types utilisés pour annoter le texte. Il permet d'appliquer une requête saisie par l'utilisateur dans le langage TextMarker sur le corpus d'entrée et d'afficher les concordances trouvées (cible, contextes gauche et droit).

Ce logiciel est développé par Sondes Bannour. Il est actuellement en phase de test. Sa diffusion est prévue début 2013.

Dafoe est une plateforme OpenSource d'acquisition d'ontologies à partir de textes. Elle repose sur une méthodologie voisine de Terminae mais avec une autre ambition : la plateforme Dafoe est conçue pour gérer des données volumineuses et un soin particulier a été apporté aux interfaces utilisateur.

La conception de cette plateforme a été réalisée dans le cadre du projet ANR Dafoe4App. Le LIPN a pris la responsabilité du travail de spécification et a participé aux tâches de test. Le développement de cette plateforme, qui a été confié à une société tierce, n'a pas pu être achevé avant la fin du projet. La poursuite du travail nécessite un investissement de la part des partenaires du projet qu'ils n'ont pas encore pu faire.

LIPN UIMA Platform est une plateforme d'annotation développée en utilisant le framework de développement Apache-UIMA. Cette plateforme repose sur un système de types générique qui met l'accent sur la nature des annotations à représenter (segmentation, étiquette, relation . . .) et qui supporte les annotations conflictuelles.

Actuellement, cette plateforme contient des composants permettant d'encapsuler les outils suivants : TagEN, TreeTagger, LIA tools et YaTeA. Elle est développée par Erwan Moreau et Laurent Audibert dans le cadre du programme Quaero.

1. <http://www-lipn.univ-paris13.fr/~szulman/Annotator/annotator.html>

SemEx² est une plateforme d'aide à la construction et à l'exploration d'un modèle métier à partir de textes réglementaires et procéduraux. SemEx repose sur les standards du W3C (SKOS, RDF, RDFa, SPARQL) et s'appuie sur le moteur sémantique CORESE³.

Ce logiciel est développé dans le cadre du projet ONTORULE par Abdoulaye Guissé.

Terminae⁴ est une plateforme d'aide à la construction de ressources termino-ontologiques à partir de sources textuelles. Elle met en œuvre la méthodologie Terminae qui guide l'ingénieur de la connaissance dans le travail de construction d'une ontologie. Elle s'appuie sur un niveau termino-conceptuel intermédiaire pour permettre le passage du texte et des connaissances terminologiques qui en sont extraites au niveau conceptuel (l'ontologie). Cette plateforme permet de visualiser les résultats d'outils de traitement automatique des langues (des extracteurs de termes et des outils de reconnaissance d'entités nommées, notamment) et de les utiliser pour construire des ressources termino-ontologiques (terminologies, thésaurus, ontologies) où les éléments conceptuels disposent d'un ancrage linguistique.

La plateforme est réalisée par une application eclipse. Elle supporte l'import/export en SKOS et en OWL. Dans la version actuelle, le niveau conceptuel est réalisé par le plugin Neon Toolkit 2.4⁵. Les versions disponibles permettent de travailler à partir de textes français et anglais. Terminae est développée par Sylvie Szulman.

Termometer⁶ est un utilitaire permettant d'évaluer une liste de termes au regard d'une référence. La mesure de distance terminologique sur laquelle il repose permet de tenir compte de l'aspect graduel de la pertinence terminologique dans l'évaluation des sorties d'extracteur de termes.

Termometer a été développé par Thibault Mondary et Jonathan van Puymbrouck dans le cadre du programme Quaero où il est utilisé dans des campagnes d'évaluation [CI-62]. Cet utilitaire est disponible sur SourceForge, sous licence GPLv3.

Terrier SIR est une plateforme OpenSource qui enrichit la plateforme de recherche d'information Terrier IR de fonctionnalités sémantiques modulaires. Elle favorise aussi bien l'implémentation et le test de méthodes de recherche d'informations sémantiques que leurs éventuelles mises à jour ou adaptations.

La plateforme est développée par Ines Bannour dans le cadre du projet Legilocal. Elle est en phase de test et sa diffusion est prévue fin 2012.

ValiTerms⁷ est un outil qui a pour objectif de faciliter la validation d'une liste de termes. ValiTerms propose une interface web qui prend en compte un corpus étiqueté morpho-syntaxiquement et une liste de termes. À l'aide de cette interface, l'expert est en mesure d'analyser rapidement les termes dans leur contexte (la phrase) et décider de leur validité. La gestion de session permet de reprendre un travail en cours et de stocker les résultats à toutes les étapes du travail.

ValiTerms est développé par Thibault Mondary dans le cadre du programme Quaero. Il est utilisé par des experts de l'INIST pour l'enrichissement de thésaurus et durant la phase d'adjudication des campagnes d'évaluation. Cet outil est disponible sur SourceForge, sous licence GPLv3.

2. <http://www-lipn.univ-paris13.fr/~guisse/index.php?n=Semex.Semex>

3. <http://www-sop.inria.fr/edelweiss/software/corese>

4. <http://lipn.fr/terminae/index.php/Download>

5. http://neon-toolkit.org/wiki/Main_Page

6. <http://sourceforge.net/projects/termometerxd/>

7. <http://sourceforge.net/projects/valiterms>

3.4 Dissémination de la recherche

YaTeA (Yet another Term extrActor)⁸ identifie et extrait des groupes nominaux pouvant être des termes, *i.e.* des termes candidats. Chaque terme candidat est analysé syntaxiquement pour faire apparaître sa structure sous la forme de relations tête-modifieur. L'extraction de termes candidats repose sur une stratégie hybride par laquelle l'extraction à partir de patrons syntaxiques construits manuellement peut être guidée et corrigée à l'aide de ressources terminologiques existantes (désambiguïsation exogène). Ces termes attestés aident au repérage des groupes, à leur analyse syntaxique et participent également à l'extraction des termes candidats.

YaTeA, qui a été développé par Thierry Hamon et Sophie Aubin, est disponible sous licence GPL sous la forme d'un module CPAN.

3.4 Dissémination de la recherche

- Adeline Nazarenko : « Traçabilité et interprétabilité : les défis de l'analyse sémantique de contenus », exposé lors des *Déjeuners de la Technologie* « *Tatouage, traçabilité et stockage de contenus numériques* ». Organisé par Plaine Commune, le Centre Francilien de l'Innovation et la Chambre de Commerce et d'Industrie de Paris Seine-Saint-Denis, Saint-Denis, 29 mai 2012.
- Adeline Nazarenko : « Ressources pour l'annotation sémantique », présentation invitée à la Table ronde Quaero, *Second séminaire Veille stratégique, scientifique et technologique (VSST 2009)*, mars 2009.
- Adeline Nazarenko, Valentina Dragos : « Towards an electronic encyclopaedia of organic chemistry : Knowledge engineering perspectives in the EnCOre project », Séminaire auprès des partenaires chimistes du projet EnCOre, Spa, Belgique, 1er oct. 2008.
- Adeline Nazarenko : « Les moteurs de recherche, comment ça marche et à quoi ça sert ? », mini-conférence Savantes Banlieue, Université Paris 13, Octobre 2007 et 2008 .
- Adeline Nazarenko : « La fouille de textes, méthodes et enjeux », exposé lors des *Déjeuners de la Technologie* « *Convergences technologiques - gestion de documents électroniques : défis et innovations* » organisé par Plaine Commune et Plaine de France, Saint-Denis, 23 mars 2007.

**Production
scientifique,
valorisation,
dissémination**

8. <http://search.cpan.org/~thhamon/Lingua-YaTeA/>

Chapitre 4

Rayonnement

Le rayonnement de l'équipe RCLN dans le monde académique se traduit par une forte implication au niveau national dans l'animation des communautés scientifiques auxquelles elle se rattache : participation au comité de rédaction de la revue *Traitement Automatique des Langues*, organisation de conférences et présidences de comités de programme, participation à des comités d'évaluation (ANR, AERES, etc.).

L'équipe RCLN jouit également d'une bonne crédibilité dans le dialogue interdisciplinaire comme le montrent le montage du projet de LabEx EFL, le rôle d'interface avec INSHS au conseil scientifique de l'INS2I et la collaboration avec la MSH de Paris-Nord.

Les efforts faits pour améliorer la visibilité de l'équipe à l'échelle européenne et internationale (invitations de chercheurs étrangers, participation à des projets européen et transnational, plus forte participation aux conférences internationales) ont déjà permis d'augmenter l'attractivité de l'équipe en termes de recrutement et devraient porter leurs fruits dans les années à venir.

4.1 Comités d'organisation et d'édition

4.1.1 Comités éditoriaux

• Membres de comités éditoriaux

- *Traitement Automatique des Langues*, 2007-2012 ; A. Nazarenko
- *Journal of Artificial General Intelligence*, 2008-2012 ; C. Recanati
- *Collection Sciences Cognitives aux Éditions de la Maison des Sciences de l'Homme*, 2007-2012 ; D. Kayser

• Autres activités de relecture

Les membres de l'équipe ont participé à des relectures, en particulier dans des numéros spéciaux, pour les revues suivantes :

- *Information Processing and Management*
- *Engineering Applications of Artificial Intelligence*
- *Fuzzy Sets and Systems*
- *Intellectica*
- *Document Numérique*
- *Technique et Science Informatiques*
- *Revue des Nouvelles Technologies de l'Information*
- *Traitement Automatique des Langues*

4.1.2 Organisation de conférences

• Présidence de comités d'organisation

- TALN, *Conférence nationale sur le Traitement Automatique des Langues Naturelles*, juin 2009, Senlis (175 participants) ; A. Nazarenko , H.Zargayouna¹

• Autres participations à des comités d'organisation

- TALN, *Conférence nationale sur le Traitement Automatique des Langues Naturelles*, juin 2009, S. Szulman, A. Rozenknop (<http://lipn.univ-paris13.fr/taln09/>)
- IC, *23èmes Journées francophones d'Ingénierie des Connaissances*, juin 2012, Paris ; S. Ben Abbès, A. Guissé, N. Mimouni, N. Omrane
- ATALA, *Workshop et journées ATALA sur les Champs Aléatoires Conditionnels*, mars 2010, A. Rozenknop

4.1.3 Comités de programme

• Présidence de comités de programme

- EuroCogSci, *European Conference on Cognitive Science*, 2007 ; D. Kayser (co-président)²
- TALN, *Conférence nationale sur le Traitement Automatique des Langues Naturelles*, juin 2009 ; A. Nazarenko , T. Poibeau³
- RECITAL, *Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, 2009 ; T. Mondary (président), A. Bossard & T. Hamon (co-présidents)⁴
- TIA, *Terminologie et Intelligence Artificielle*, 2009, Toulouse ; S. Szulman (co-présidente)⁵

1. <http://lipn.univ-paris13.fr/taln09/>

2. <http://conferences.phs.uoa.gr/EuroCogSci07/>

3. <http://lipn.univ-paris13.fr/taln09/>

4. <http://lipn.univ-paris13.fr/taln09/index.php?conf=RECITAL>

5. <http://tia2011.crim.fr/Home.html>

- IC, 23èmes Journées francophones d'Ingénierie des Connaissances, juin 2012, Paris ; S. Szulman⁶

- **Participation à des comités de programme de conférences internationales importantes**

- EKAW, *International Conference on Knowledge Engineering and Knowledge Management*
- EMNLP, *Conference on Empirical Methods in Natural Language Processing*
- EuroCogSci, *European Conference on Cognitive Science*
- ICAI, *International Conference on Artificial Intelligence*
- ICTAI, *International Conference on Tools with Artificial Intelligence*
- IJCAI, *International Joint Conference on Artificial Intelligence*
- LREC, *Language Resources and Evaluation Conference*
- SAAKM, *Semantic Authoring, Annotation and Knowledge Markup*
- SARA, *Symposium on Abstraction, Reformulation and Approximation*
- SMBM, *International Symposium on Semantic Mining in Biomedicine*
- TKE, *Terminology and Knowledge Engineering*

- **Autres comités de programme**

Les membres de l'équipe RCLN participent également aux comités de programme des conférences de leur domaine de recherche : notamment CORIA, IC, LREC, RAPC, TALN, TIA.

4.2 Expertises

4.2.1 Comités d'évaluation et conseils

- **Comité d'évaluation AERES**

- 4 Equipes d'accueil de linguistique (Université Paris 3), février 2008 ; A. Nazarenko
- Lattice, UMR 8094 (CNRS, Ecole Normale Supérieure & Université Paris 3), février 2009 ; A. Nazarenko
- Structure fédérative Agorantic, (Université d'Avignon et des Pays de Vaucluse), janvier 2012 ; A. Nazarenko (présidente)
- Masters de l'Université de Lorraine dans le domaine Sciences, Technologies, Santé, mars 2012 ; A. Nazarenko
- Masters, nov. déc. 2009 ; D. Kayser (président)
- CEDRIC, EA 1395 (CNAM), mai 2010 ; D. Kayser (président)
- Modélisation Information et Systèmes (M.I.S.), EA 4290 (Université de Picardie Jules Verne), février 2011 ; D. Kayser (président)

- **Autres comités d'évaluation**

- Comité d'évaluation pour les appels à projets ANR Blanc et Jeunes Chercheuses et Jeunes Chercheurs en Sciences et Technologie de l'Information et de la Communication, 2007-2009 ; D. Kayser (président)
- Comité scientifique disciplinaire STIC du programme ANR Blanc et Jeunes Chercheurs, mars 2007 - juillet 2009 ; D. Kayser (président)
- Comité d'évaluation pour les appels à projets ANR "Contenu et Interaction", 2008-2009 ; A. Nazarenko
- Commission thématique "Ingénierie des Connaissances" du pôle de compétitivité Cap Digital-Paris Région, 2005-2008 ; A. Nazarenko

6. <http://ic2012.crc.jussieu.fr/>

4.3 Invitations

4.2.2 Autres participations à des conseils

- Direction adjointe du LABEX « Fondements empiriques de la linguistique » (Labex EFL), et responsable de l'axe "Analyse sémantique computationnelle", 2011-12 ; A. Nazarenko
- Conseil scientifique du LABEX Empirical Foundations of Linguistics (EFL) ; A. Rozenknop
- Conseil scientifique du GDR Psycho Ergo, 2008-2011 ; D. Kayser
- Comité scientifique disciplinaire STIC de l'ANR (2007-09), D. Kayser (président)
- Conseil d'administration du pôle de compétitivité Cap Digital (2006-12) ; A. Nazarenko

4.2.3 Autres expertises de projets

- Programme ANR Corpus, 2007 ; A. Nazarenko, F. Gayral
- Programme ANR TECSAN, 2007 ; A. Nazarenko
- Programme ANR Blanc, 2007-2009 ; A. Nazarenko, F. Levy
- Programme ANR Jeunes Chercheuses et Jeunes Chercheurs, 2007 ; A. Nazarenko
- Programme ANR Emergence, 2010 ; H. Zargayouna
- Projets de chercheurs canadiens (NSERC, 2009) et suisse (2011) ; A. Nazarenko
- Projets-contrats de la région de la Basse-Normandie (2) ; A. Nazarenko
- Projet DIGITEO, 2012 ; A. Nazarenko
- ANRT bourses CIFRE, 2008 ; H. Zargayouna

4.3 Invitations

4.3.1 Invitations des membres du LIPN

• Conférences invitées, tutoriels, animation de tables rondes

- Haïfa Zargayouna : Animation de la table ronde *Quelle évaluation pour la Recherche d'Information Sémantique*, Atelier Recherche d'Information Sémantique (RISE@CORIA), mars 2011.
- Haïfa Zargayouna : conférence invitée *Le Web entre norme et anarchie*, Journées d'étude *Le web a-t-il un sens ?* du groupe TICIS-SFSIC, Université Paris 13, décembre 2010.
- Adeline Nazarenko : co-organisation d'un tutoriel *How to Integrate Ontologies and Rules ?*, associé à la 24th AAAI Conference on Artificial Intelligence, Atlanta. Présentateurs : Thomas Eiter (TU Vienna), Luis Polo (Fundación CTIC), Adeline Nazarenko (Université Paris 13), Juillet 2010.
- Adeline Nazarenko : co-organisation d'un tutoriel *Integrating Ontologies and Rules in a Semantic Business : From Policies to Operation*, associé à la 7th Extended Semantic Web Conference (ESWC 2010) in Heraklion, Crete (Greece). Présentateurs : Thomas Eiter (TU Vienna), Stijn Heymans (TU Vienna), Luis Polo (Fundación CTIC), Adeline Nazarenko (Université Paris 13), mai 2010.

• Séminaires

- François Lévy, Adeline Nazarenko, Abdoulaye Guissé : *D'où viennent les règles des systèmes experts (d'aujourd'hui) ? le rôle du texte*, Exposé sur invitation aux journées TAL et IA, 12 mars 2012.
- Adeline Nazarenko, Nouha Omrane, Sylvie Szulman : *Indices textuels et connaissances de domaine : apport du TAL à l'ICT*, Exposé sur invitation aux journées TAL et IA, 12 mars 2012.

- Sylvie Szulman : *Méthodes et outils de construction d'ontologies à partir de textes*, Séminaire invité par le laboratoire I3S (UMR 6070) et l'équipe Edelweiss de l'Inria Sophia Antipolis, Novembre 2011.
- Daniel Kayser : *Quelques réflexions sur le Traitement Automatique des Langues Naturelles*, Séminaire à l'Association Européenne Interdisciplinaire, Paris, Décembre 2010.
- Daniel Kayser : *Un malentendu fécond entre l'Intelligence Artificielle et la Psychologie Cognitive : La question de la représentation*, Journée à la mémoire de J.F Le Ny, Université Paris-Sorbonne, Septembre 2007.

Les membres de l'équipe RCLN ont par ailleurs donné une trentaine de séminaires dans des laboratoires variés.

4.3.2 Invitations à Paris 13

Prénom	Nom	Origine	Période	Financement
Anna	Korhonen	Royal Society University Research Fellow	mai et juillet 2007	maître de conférence invitée
Aldo	Gangemi	STLab & CNR (IT) Senior researcher	juin 2011	professeur invité
Aldo	Gangemi	STLab & CNR (IT) Senior researcher	juin 2012	professeur invité
Presutti	Valentina	STLab & CNR (IT) Researcher	sept-déc. 2012	chercheur invité (CNRS)

Chapitre 5

Formation à la recherche

L'équipe RCLN a maintenu une politique active de formation de nombreux jeunes chercheurs à travers la création d'une spécialité de master *Ingénierie des Textes et Contenus Numériques* et de nombreux cours dans différents masters, avec l'organisation de tutoriels dans de grandes conférences internationales, par l'accueil et l'encadrement de stagiaires de master 2 et de doctorants.

Le flux d'étudiants ne permettant pas d'assurer des effectifs suffisants pour 3 spécialités de master, il a été décidé de fermer la spécialité « Ingénierie des textes et contenus numériques » qui offrait cependant aux étudiants à la fois des débouchés professionnels et en recherche. L'équipe RCLN prévoit d'entamer en 2013 des discussions avec ses partenaires du LabEx EFL et notamment Alpage pour ouvrir une nouvelle spécialité, plus orientée recherche, en s'appuyant sur les forces du LabEx. L'articulation de cette nouvelle spécialité avec celles qui existent reste à définir.

5.1 Masters, Ecoles doctorale et post-doctorale

L'équipe RCLN prend une part active à la formation à la recherche des étudiants pour attirer de nouveaux doctorants dans l'équipe mais aussi pour assurer le rayonnement des thématiques de recherche propres à l'équipe. Qu'ils soient en poste à l'IUT ou à l'IG, la plupart des membres de l'équipe assurent couramment des cours dans le master

5.2 Thèses et Habilitations

d'informatique de l'université Paris 13, principalement dans les spécialités MICR (2007-09) et ITCN (2009-11) coordonnées par Adeline Nazarenko mais aussi dans la spécialité EID et dans le tronc commun de la 1ère et 2ème année du master d'informatique. Les membres de l'équipe apportent également leur expertise à d'autres masters.

L'équipe RCLN recrute ses futurs doctorants naturellement dans les masters ci-dessus où elle intervient mais elle recrute aussi des doctorants en provenance d'autres masters, en général après les avoir mis à l'épreuve au cours d'un stage de recherche préalable.

• Cours assurés dans le master d'informatique de l'université Paris 13

- *Outils de Manipulation de Données Textuelles (OMDT)*, ITCN ; L. Audibert.
- *Ingénierie des connaissances*, M1 ; S. Szulman, A. Nazarenko, D. Després
- *Web Sémantique et IA*, M1 ; S. Després
- *Ingénierie des connaissances textuelles*, MICR & ITCN ; A. Nazarenko, S. Szulman, S. Després, T. Hamon
- *Représentation des Connaissances*, M2 ; D. Kayser
- *Raisonnement sur l'incertain et la causalité*, MICR ; D. Kayser
- *Fondements de la représentation des connaissances*, MICR ; F. Lévy
- *Méthodologie de la recherche*, EID, MICR & ITCN ; A. Nazarenko
- *Métiers de l'informatique* M2 ; A. Nazarenko
- *Fouille de données textuelles*, EID, MICR & ITCN ; A. Nazarenko
- *Extraction d'information*, MICR ; T. Poibeau
- *Interaction Homme/Machine*, M2 ; C. Recanati
- *Traitement Automatique du Langage*, MICR ; A. Rozenknop
- *Recherche et Extraction d'Information*, MICR ; A. Rozenknop
- *Modèle de Langages et Analyse Syntaxique*, ITCN ; A. Rozenknop
- *Méthodes d'accès à l'information*, ITCN ; H. Zargayouna

• Cours assurés dans d'autres masters

- Master 2 professionnel spécialité ISV, Université Paris 5 : *Gestion des connaissances* ; S. Després :
- Master 2 Informatique linguistique Pro, Université Paris 7 : *Ingénierie des ontologies* ; S. Després
- Master 2 métiers du livre, Pôle des métiers du livre, Université Paris X : *Document numérique* ; A. Nazarenko, H. Zargayouna
- *Représentation des connaissances*, Master Ingénierie multilingue de l'INALCO ; T. Poibeau
- Master IFI, filière KIS - Université De Nice Sophia-Antipolis : *Construction de ressources sémantiques à partir de textes* ; S. Szulman

5.2 Thèses et Habilitations

5.2.1 Thèses en cours

Nom Prénom	1ère insc.	Financement
Titre	Directeur	Co-encadrant
BANNOUR Inès	01/10/12	Alloc (12-15)
<i>Accès sémantique à l'information textuelle en domaine ouvert</i>	A. Nazarenko	H. Zargayouna
MIMOUNI - MESSAI Nada	01/02/11	FUI Legilocal
<i>Modèle documentaire pour le web sémantique juridique</i>	A. Nazarenko	S. Salotti

5.2 Thèses et Habilitations

BANNOUR Sondes <i>Apprentissage interactif de règles pour l'extraction d'informations dans des documents</i>	01/10/10 H. Soldano	Programme Quaero L. Audibert
OMRANE Nouha <i>Construction d'ontologies métier à partir de textes réglementaires</i>	10/01/09 A. Nazarenko	CEE Ontorule S. Szulman
VAN PUymbrouck Jonathan <i>Exploitation de folksonomies pour la recherche d'information</i>	10/01/09 A. Nazarenko	Alloc (09-12) A. Rozenknop
BEN ABBES Sarah <i>Construction d'une cartographie de domaine à partir de ressources sémantiques hétérogènes</i>	10/09/09 A. Nazarenko	Programme Quaero H. Zargayouna
GUISSE Abdoulaye <i>Formalisme et méthodes pour l'annotation sémantique automatique de documents</i>	10/09/09 F. Lévy	Alloc (09-12)
FORT Karèn <i>Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus</i>	01/12/08 A. Nazarenko	Ing. INIST (Programme Quaero)

5.2.2 Thèses soutenues

(depuis 2007)

Nom Prénom Titre Jury : Jury	Situation ultérieure	Soutenance Directeur	Mention Co-encadrant
MONDARY Thibault <i>Construction d'ontologies à partir de textes. L'apport de l'analyse de concepts formels</i> Jury : Nathalie Aussenac-Gilles, Gilles Kassel (rapp), Nathalie Pernelle, Yannick Toussaint, Pierre Zweigenbaum (rapp)	Post-doc	27/05/11 A. Nazarenko	S. Després
BOUTOUHAMI Sara <i>Un système de génération de descriptions argumentées</i> Jury : Philippe Dague (pdt), Patrick Saint-Dizier (rapp), Jérôme Lang (rapp), Aïcha Mokhtari, Adeline Nazarenko		16/11/10 D. Kayser	
HAMON Olivier <i>Vers une architecture générique et pérenne pour l'évaluation en traitement automatique des langues : spécifications, méthodologies et mesures</i> Jury : Mohand Boughanem (rapp), Khalid Choukri (ELDA), Anthony Hartley, Daniel Kayser, Joseph Mariani, François Yvon (rapp)	Ingénieur	06/12/10 A. Nazarenko	
CAILLIAU Frédéric <i>Des ressources aux traitements linguistiques : le rôle d'une architecture linguistique</i> Jury : Claude de Lopy, Patrice Bellot (rapp), Olivier Gaunet, Nabil Hathout (rapp), François Lévy, Emmanuel Morin	Ingénieur	09/12/10 A. Nazarenko	

Formation à la recherche

5.3 Stages de master, post-doctorats

MESSIANT Cédric	Ingénieur	05/11/10	
<i>Acquisition automatique de schémas de sous-catégorisation à partir de corpus bruts</i>			
<i>Jury : Eric Laporte (pdt), Nuria Bel (rapp), Alexis Nasr (rapp), Antoine Rozenknop, Jacques Blanc-Talon</i>			
BOSSARD Aurélien	Post-doc Orange Labs	12/07/10	
<i>Contribution au résumé automatique multi-documents</i>			
<i>Jury : Guy Lapalme (UdM), Juan-Manuel Torres Moreno, Anne Vilnat, Céline Rouveiro</i>			
WEISSENBACHER Davy	Post-doc Manchester	20/11/08	TH
<i>Influence des annotations imparfaites sur les systèmes de Traitement Automatique des Langues. Un cadre applicatif : la résolution de l'anaphore pronominale</i>			
<i>Jury : Jean-Gabriel Ganascia (pdt), Sophia Ananiadou, Bernard Victorri (rapp), Philippe Leray (rapp)</i>			
BOUFFIER Amanda	IR Temis	16/10/08	THF
<i>Techniques d'apprentissage par explication et par détection de similarités</i>			
<i>Jury : Marie-Paule Pery-Woodley (rapp), Patrice Enjalbert (rapp), Jean-Pierre Desclés, Alain Venot</i>			
BOURGOIS Laure	CDD Sénégal	19/06/07	TH
<i>Représentation et comparaison de Web services complexes avec des logiques dynamiques</i>			
<i>Jury : Philippe Dague (rapp), Christophe Fouqueré (pdt), Laurent Hénocque, Farouk Toumani (rap)</i>			
NOUIOUA Farid	Post-doc LRI	25/04/07	TH
<i>Extraction et utilisation de normes pour le raisonnement causal dans un corpus textuel</i>			
<i>Jury : Gérard Sabah (rapp), Pierre Marquis (rapp), Jean-Paul Haton, Pascal Nicolas, A. Nazarenko</i>			

5.2.3 Habilitations à Diriger des Recherches soutenues

(depuis 2007)

Nom	Prénom	Soutenance
	<i>Titre</i>	
	<i>Jury :</i>	
Poibeau Thierry		26/11/08
<i>Des mots au texte. Analyse sémantique pour l'accès à l'information</i>		
<i>Jury : Catherine Fuchs (rapp), Benoît Habert (rapp), Maria Teresa Paziienza (rapp), Michel Denis, Daniel Kayser, Adeline Nazarenko, Pierre Zweigenbaum</i>		

5.3 Stages de master, post-doctorats

5.3.1 Postdoctorats

- 2006-07 : Haïfa Zargayouna, postdoc sur le projet *Infom@gic*, Conception d'un système d'indexation fine de textes (resp. A. Nazarenko).

- 2008 : Valentina Dragos, postdoc dans le cadre de son contrat d'ATER à l'Université Paris 13, *Conception d'une architecture de système d'accès à l'information scientifique qualifiée en chimie organique* (resp. A. Nazarenko & A. Krief, Université de Namur, Belgique).
- 2007-08 : Marie-Paule Jacques, postdoc sur le projet ANR *TextCoop*, *Méthodes d'analyse des types de textes : état de l'art et expérimentations* (resp. F. Gayral).
- 2008-09 : Erwan Moreau, postdoc sur le projet ANR *Crotal*, *Conception d'une plateforme d'annotation avec UIMA* (resp. A. Rozenknop, T. Poibeau).
- 2010 : Erwan Moreau, postdoc dans le cadre du programme *Quaero*, *Conception d'une plateforme d'annotation avec UIMA* (resp. L. Audibert).
- 2010 : Rim Bentebibel, postdoc sur le projet ANR *Dafoe*, *Conception d'une méthode d'extraction de relations de domaine à partir de textes* (resp. S. Szulman, A. Nazarenko).
- 2008-2011 : Yue Ma, postdoc dans le cadre du programme *Quaero*, *Proposition d'un formalisme unifiant la représentation des ressources terminologiques et ontologiques* (resp. L. Audibert, A. Nazarenko).
- 2011-2012 : Thibault Mondary, postdoc dans le cadre du programme *Quaero*, *Conception et gestion de campagnes d'évaluation* (resp. H. Zargayouna, A. Nazarenko).
- 2011-2012 : Yue Ma, postdoc dans le cadre du programme *Quaero*, *Annotation sémantique* (resp. F. Lévy, A. Nazarenko).

5.3.2 Stages de master

Sauf mention contraire, la durée des stages de master est de 6 mois. L'université d'origine n'est mentionnée que lorsque les étudiants ne proviennent pas de l'Université Paris 13.

- 2007 (T. Hamon) : Neeven Cadapen, *Filtrage de termes extraits d'un corpus*.
- 2007 (D. Weissenbacher, A. Nazarenko) : Ayat-Allah Bouchouareb, *Fouille de texte pour l'enrichissement d'une base de données de noms de gènes*.
- 2007 (D. Kayser) : Nicolas Kamennoff, *Analyse sémantique, modélisation et visualisation : constats d'accidents de la route*.
- 2007 (H. Zargayouna) : Myriam Rakho, Stage master de linguistique Informatique (Université Paris 7) 1ère année (4 mois), *Générer des index de fin de livre et alignement de renvois vers les textes (de XML à HTML)*.
- 2008 (A. Nazarenko) : Mourad El Moueddeb, *Définition d'un protocole d'évaluation pour les outils d'analyse terminologique*.
- 2008 (M. Généreux) : Laurent Ly, *Analyse textuelle pour la détection de niveaux de crise*.
- 2009 (L. Audibert, A. Nazarenko) : Sondes Bannour, Master 2 professionnel, Ecole Nationale des Sciences de l'Informatique (ENSI, Tunisie), *Mise en oeuvre d'un environnement de développement UIMA dans le cadre d'une chaîne d'annotation linguistique*.
- 2009 (F. Lévy, A. Nazarenko) : Abdoulaye Guissé, Master 2 recherche, Université Gaston Berger (Saint Louis du Sénégal), *Annotation sémantique de documents : développement d'un outil d'indexation et de navigation pour la révision des règles métiers*.
- 2009 (C. Recanati) : Adel ben Mertah, *Langage pseudo naturel comportant des expressions spatiales (et/ou temporelles) pour un dialogue homme-machine multi-modal*.
- 2009 (A. Rozenknop) : Djamel Fali, *Optimisations pour l'apprentissage statistique sur corpus arborés*,
- 2009 (S. Szulman, A. Nazarenko) : Nouha Omrane, Master 2 Informatique, Systèmes intelligents (ISI), Université Paris Dauphine, *méthodes de construction d'une ontologie de domaine à partir de textes réglementaires*.

5.3 Stages de master, post-doctorats

- 2009 (H. Zargayouna) : Sarra Ben Abbès, Master 2 Informatique, Systèmes intelligents (ISI), Université Paris Dauphine, *Évaluation des méthodes d'acquisition d'ontologies*.
- 2010 (L. Audibert) : Sondes Bannour, Master 2 recherche, Ecole Nationale des Sciences de l'Informatique (ENSI, Tunisie), *Mesure de similarité distributionnelle entre termes*.
- 2010 (Y. Ma) : Sudeep Ghimire, stage de Master European Masters Program in Computational Logic, Freie Universität Bozen et Technische Universität Dresden, *Reasoning with Text Annotations*.
- 2010 (S. Salotti, H. Zargayouna) : Imen Ouardani, Master 2 Informatique, Systèmes intelligents (ISI), Université Paris Dauphine, *Indexation et recherche d'information sémantique*.
- 2010 (S. Szulman, A. Nazarenko) : Oumy Seye, Master 2, Université Gaston Berger de Saint Louis Sénégal, *Réalisation et évaluation d'un système d'extraction de relations sémantiques*. 2011 (H. Zargayouna, S. Ben Abbès) : Daouda Sangaré, *Méta-évaluation de l'acquisition de classes sémantiques de mots*.
- 2012 (F. Lévy) : Leidiana Martins, stage de Master Sciences du langage, INALCO (stage à l'IGN), *Extraction de règles concernant les bâtiments d'un corpus de plans locaux d'urbanisme (PLU)*.
- 2012 (H. Zargayouna) : Ines Bannour, Master 2 Recherche, École Nationale des Sciences de l'Informatique (ENSI) Tunisie, *Méthodes d'indexation et de recherche d'information pour le Web Sémantique*.

5.3.3 Autres stages et encadrements

- 2008 (T. Hamon) : encadrement de Nicolas Beauger, stage d'ingénieur (3 mois), *Mise en oeuvre d'outils de traitement automatique des langues*.
- 2008 (L. Audibert) : encadrement du mémoire de fin d'étude de Simon Cacheux, ENS Louis Lumière, *Une classification des sons en sonothèque*.
- 2011 (H. Zargayouna) : encadrement de Ines Bannour, École Nationale des Sciences de l'Informatique (ENSI, Tunisie), Stage ingénieur (4 mois), *Étude comparative de méthodes de recherche d'information sémantique*.

5.3.4 Jurys de thèse et d'HDR (extérieurs à l'Université Paris 13)

Sur la période les membres de l'équipe RCLN ont participé à 17 jurys de thèses extérieures à l'Université Paris 13, dont 9 comme rapporteur. A noter également 2 jurys de thèses à l'étranger.

Ils ont également participé à 6 jurys d'habilitation.

Chapitre 6 Publications

Les références comportant « * » sont acceptées et à paraître.

Signification des soulignements :

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

Sommaire

Articles dans des revues avec comité de lecture et chapitres de livres	
(ACL)	56
Communications avec actes (ACT)	59
Direction d'ouvrages (DO)	70
Logiciels et Brevets (LO)	71
Thèses et Habilitations (TH)	72
Divers (Div)	73

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

Articles dans des revues internationales avec comité de lecture sélectif

- 2012** [RI-26*] B. Crabbé, D. Duchier, C. Gardent, J. Le Roux et Y. Parmentier. XMG : eXtensible Meta-Grammar. *Computational Linguistics*, pages 1–38, août 2012.
- [RI-25*] F. Maier, Y. Ma et P. Hitzler. Paraconsistent OWL and Related Logics. *Semantic Web journal*, mai 2012.
- [RI-24] A. Rozenknop, R. Wolfler Calvo, L. Alfandari, D. Chemla et L. Létocart. Solving electricity production planning by column generation. *Journal of Scheduling*, 15(6) :1–20, septembre 2012.
- [RI-23] D. Weissenbacher et A. Nazarenko. Comprendre les effets des erreurs d’annotations des plates-formes de TAL. *Traitement Automatique des Langues (TAL)*, 52 :161–185, janvier 2012.
- [RI-22] L. Zhou, H. Huang, G. Qi, Y. Ma, Z. Huang et Y. Qu. Paraconsistent query answering over DL-Lite ontologies. *Web Intelligence and Agent Systems*, 10(1) :19–31, mars 2012.
- 2011** [RI-21] K. Fort, G. Adda et K. Bretonnel Cohen. Amazon Mechanical Turk : Gold Mine or Coal Mine? *Computational Linguistics*, pages 413–420, 2011. Last words.
- 2010** [RI-20] A. Bossard, M. Généreux et T. Poibeau. Résumé automatique de textes d’opinion. *Traitement Automatique des Langues (TAL)*, 51(3) :47–73, 2010.
- [RI-19] S. Boutouhami et D. Kayser. Strategies to defend a protagonist of an event. *International Journal on Artificial Intelligence Tools*, vol. 19(n°4) :439–464, août 2010.
- [RI-18] F. Cailliau et A. Cavet. Analyse des sentiments et transcription automatique : modélisation du déroulement de conversations téléphonique. *Traitement Automatique des Langues (TAL)*, pages 131–154, 2010.
- [RI-17] D. Kayser. The place of Logic in reasoning. *Logica Universalis*, vol. 4(n°2) :225–239, novembre 2010.
- [RI-16] Y. Ma, G. Qi et P. Hitzler. Computing Inconsistency Measure based on Paraconsistent Semantics. *Journal of Logic and Computation*, 21(6) :1257–1281, décembre 2010.
- [RI-15] Y. Ma, G. Qi, G. Xiao, P. Hitzler et Z. Lin. Computational Complexity and Anytime Algorithm for Inconsistency Measurement. *International Journal of Software and Informatics*, 4(1) :3–21, mars 2010.
- [RI-14] C. Messiant, K. Gábor et T. Poibeau. Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. *Traitement Automatique des Langues (TAL)*, 51(1) :65–96, novembre 2010.
- [RI-13] F. Zehraoui, R. Kanawati et S. Salotti. Hybrid neural network and case based reasoning system for Web user behavior clustering and classification. *International Journal of Hybrid Intelligent Systems*, 7(3) :171–186, août 2010.
- 2009** [RI-12] A. Bouffier. A textual approach for the analysis of health practice guidelines. *Traitement Automatique des Langues (TAL)*, 50(1) :35–59, 2009.
- [RI-11] N. Grabar, P.-C. Varoutas, P. Rizand, A. Livartowski et T. Hamon. Automatic acquisition of synonym resources and assessment of their impact on the enhanced search in EHRs. *Methods of Information in Medicine*, 48(2) :149–154, 2009.
- [RI-10] D. Kayser et F. Nouioua. From the Description of an Accident to its Causes. *Artificial Intelligence*, 173(12-13) :1154–1193, 2009.
- [RI-9] A. Nazarenko, H. Zargayouna, O. Hamon et J. Van Puymbrouck. Evaluation des outils terminologiques : enjeux, difficultés et propositions. *Traitement Automatique des Langues (TAL)*, 50(1 varia) :257–281, 2009.
- [RI-8] T. Poibeau. Boosting Robustness of a Named Entity Recognizer. *International Journal of Semantic Computing*, 3(1) :1–14, 2009.
- [RI-7] T. Poibeau et D. Dutoit. Automatic extraction of paraphrastic phrases from small size corpora. *Linguisticae Investigationes*, 32(1) :77–98, 2009.
- 2008** [RI-6] O. Hamon, P. Paroubek et D. Mostefa. SEWS : un serveur d’évaluation orienté Web pour la syntaxe. *Traitement Automatique des Langues (TAL)*, 49(2) :247–270, 2008.

- [RI-5] T. Hamon et A. Nazarenko. Le développement d'une plate-forme pour l'annotation spécialisée de documents Web : retour d'expérience. *Traitement Automatique des Langues (TAL)*, 49(2) :127–154, 2008.
- [RI-4] T. Poibeau et A. Condamines. Linguistique et accès automatisé à l'information : un bilan. *Revue française de linguistique appliquée*, 13(1) :5–8, 2008.
- 2007** [RI-3] S. Després et S. Szulman. Merging of Legal Micro-Ontologies from European Directives. *Journal of Artificial Intelligence and Law*, 15(2) :187–200, février 2007.
- [RI-2] K. Gollub, T. Hamon et A. Ardö. Automated classification of textual documents based on a controlled vocabulary in engineering. *Knowledge Organization*, 34(4) :247–263, novembre 2007.
- [RI-1] T. Poibeau. Semantic annotation : Mapping text to ontologies. *International Journal of Metadata Semantics and Ontologies*, 2(2), 2007.

Articles dans des revues nationales ou internationales

- 2010** [RE-5] F. Lévy, A. Nazarenko et A. Guissé. Annotation, indexation et parcours de documents numériques. *Revue des Sciences et Technologies de l'Information (Série IDocument Numérique)*, 13(3/2010) :121–152, décembre 2010.
- 2008** [RE-4] S. Boutouhami et D. Kayser. Vers la construction de descriptions argumentées d'un accident de la route : analyse de diverses stratégies argumentatives. *Revue de sciences du langage : COgnition, REprésentation, LAngage (CORELA 2008)*, vol. 6(n°1), juin 2008. édition électronique.
- [RE-3] S. Després et H. Richy. Métadonnées, ontologies et documents numériques. *Revue des Techniques de l'Ingénieur*, pages 1–19, 2008.
- 2007** [RE-2] S. Aubin, J. Derivière, T. Hamon, A. Nazarenko, T. Poibeau et D. Weissenbacher. Une infrastructure pour l'annotation linguistique de documents issus du web : le projet ALVIS. *Revue des Nouvelles Technologies de l'Information*, pages 55–72, 2007.
- [RE-1] D. Weissenbacher. Les réseaux bayésiens : un formalisme adapté au traitement automatique des langues? *Revue d'Intelligence Artificielle*, 21(3) :371–389, juin 2007.

Chapitres de livre

- 2011** [CL-8] C. Garbay et D. Kayser. Introduction. C. Garbay et D. Kayser, éditeurs, *Informatique et sciences cognitives : Influences ou confluence?*, pages 11–19. Editions de la MSH, juin 2011.
- 2010** [CL-7] F. Hamdi, B. Safar, C. Reynaud et H. Zargayouna. Alignment-based Partitioning of Large-scale Ontologies. F. Guillet, G. Ritschard, D. Zighed et H. Briand, éditeurs, *Advances in Knowledge Discovery And Management*, volume 292, pages 251–269. Springer, 2010.
- 2009** [CL-6] C. Nédellec, A. Nazarenko et R. Bossy. Information Extraction. S. Staab et R. Studer, éditeurs, *Handbook on Ontologies*, Springer Series on Handbooks in Information Systems, pages 663–686. Springer Verlag, 2009. 2nd revised edition.
- 2008** [CL-5] N. Aussenac-Gilles, S. Després et S. Szulman. The Terminae Method and Platform for Ontology Engineering from Texts. *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, pages 192–223. IOS press, février 2008.
- 2007** [CL-4] A. Nazarenko et T. Aït El Mekki. Building back-of-the-book indexes. F. Ibekwe-SanJuan, A. Condamines et M. T. Cabré, éditeurs, *Application-Driven Terminology Engineering*, Current Topics 2, pages 179–201. John Benjamins, 2007.
- [CL-3] T. Poibeau et A. Vilnat. Traitement automatique des langues et Systèmes de Question-Réponses (SQR). B. Grau et J.-P. Chevallet, éditeurs, *La recherche d'informations précises. Traitement automatique de la langue, apprentissage et connaissances pour les systèmes de question-réponse*, pages 105–130. Hermès-Lavoisier, 2007.
- [CL-2] M. Rajman, J.-C. Chappelier et A. Rozenknop. Syntactic Processing. M. Rajman et V. Palola, éditeurs, *Speech and Language Engineering*, pages 177–204. Presses Polytechniques et Universitaires Romandes (CH)/ Marcel Dekker (US), 2007.
- [CL-1] F. Zehraoui, R. Kanawati et S. Salotti. CASEP2 : système hybride pour le traitement de séquences de navigation sur le web. J. Renaud, B. Chebel, B. Fuchs et J. Lieber, éditeurs, *Raisonnement à partir de cas 2*, Traité IC2, pages 151–176. Hermès, 2007.

Autres communications sur invitation

- 2012** [AI-2] K. Fort. Annotation manuelle de corpus : mais de quoi parle-t-on ? Journée « Annotation collaborative de corpus », MSH Nancy, France, 2012. Présentation invitée.
- 2010** [AI-1] H. Zargayouna. Le Web entre norme et anarchie. *Journées d'étude « le web a-t-il un sens ? »*, Villetaneuse, France, décembre 2010.

Communications avec actes (ACT)

Communications dans des conférences internationales avec comité de lecture

- 2012**
- [CI-67] K. Fort et V. Claveau. Annotating Football Matches : Influence of the Source Medium on Manual Annotation. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2567–2572, Istanbul, Turquie, mai 2012.
 - [CI-66] K. Fort, C. François, O. Galibert et M. Ghribi. Analyzing the Impact of Prevalence on the Evaluation of a Manual Annotation Campaign. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 1474–1480, Istanbul, Turquie, mai 2012.
 - [CI-65*] K. Fort, A. Nazarenko et S. Rosset. Modeling the Complexity of Manual Annotation Tasks : a Grid of Analysis. *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, décembre 2012.
 - [CI-64] A. Guissé, F. Lévy et A. Nazarenko. From regulatory texts to BRMS : how to guide the acquisition of business rules? *Proceedings of the 6th International Symposium On Rules : Research Based And Industry Focused (RuleML 2012)*, pages 77–91, Montpellier, France, 2012.
 - [CI-63] S. A. Mirroshandel, A. Nasr et J. Le Roux. Semi-supervised Dependency Parsing using Lexical Affinities. *Proceedings of the 50th annual meeting of the Association for Computational Linguistics (ACL 2012)*, pages 777–785, Corée, République De, 2012.
 - [CI-62] T. Mondary, A. Nazarenko, H. Zargayouna et S. Barreaux. The Quaero Evaluation Campaign on Term Extraction. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 663–669, Istanbul, Turquie, mai 2012.
 - [CI-61*] G. Peeters et K. Fort. Towards a (better) Definition of Annotated MIR Corpora. *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, octobre 2012.
 - [CI-60*] G. Xiao et Y. Ma. Inconsistency Measurement based on Variables in Minimal Unsatisfiable Subsets. *Proceedings of the European Conference on Artificial Intelligence 2012 (ECAI'12)*, pages 864–869, France, août 2012.
- 2011**
- [CI-59] G. Adda, B. Sagot, K. Fort et J. Mariani. Crowdsourcing for Language Resource Development : Critical Analysis of Amazon Mechanical Turk Overpowering Use. *Proceedings of the 5th Language and Technology Conference (LTC 2011)*, Poznan, Pologne, novembre 2011.
 - [CI-58] S. Ben Abbès, H. Zargayouna et A. Nazarenko. Evaluating semantic classes used for ontology building and learning from texts. *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD 2011)*, numéro 85, pages 445–448, Paris, France, mai 2011. Senart Campus of Université Paris-Est Créteil (UPEC). Quaero.
 - [CI-57] K. Fort, A. Nazarenko et C. Ris. Corpus Linguistics for the Annotation Manager. *Proceedings of the Corpus Linguistics Conference (CL 2011)*, pages 1–13, Birmingham, Royaume-Uni, novembre 2011. Quaero.
 - [CI-56] Y. Ma, F. Lévy et S. Ghimire. Reasoning with Annotations of Texts. *Proceedings of the 24th Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*, pages 192–197, États-Unis, mai 2011.
 - [CI-55] A. Nazarenko, A. Guissé, F. Lévy, N. Omrane et S. Szulman. Integrating Written Policies in Business Rule Management Systems. *Rule-Based Reasoning, Programming, and Applications, Proceedings of the RuleML 2011 Conference (Europe)*, volume 6826 de *Lecture Notes in Computer Science*, pages 99–113, Barcelona, Espagne, 2011. Springerlink.
 - [CI-54] N. Omrane, A. Nazarenko, P. Rosina, S. Szulman et C. Westphal. Lexicalized ontology for a business rules management platform : An automotive use case. Springer, éditeur, *Rule-Based Modeling and Computing in the Semantic Web (Proceedings of the RuleML 2011 - America)*, LNCS 7018, pages 179–192, Florida, États-Unis, novembre 2011.
 - [CI-53] N. Omrane, A. Nazarenko et S. Szulman. From Linguistics to Ontologies The Role of Named Entities in the Conceptualisation Process. *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD 2011)*, pages 249–254, Paris, France, octobre 2011.

- 2010**
- [CI-52] O. Hamon. Is my Judge a good One? *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010.
 - [CI-51] F. Lévy, A. Guisse, A. Nazarenko, N. Omrane et S. Szulman. An Environment for the Joint Management of Written Policies and Business Rules. E. Grégoire, éditeur, *Proceedings of the 22nd International Conference on Tools with Artificial Intelligence (ICTAI 2010)*, volume II, pages 142–149, Arras, France, octobre 2010. IEE-CPS.
 - [CI-50] Y. Ma, A. Nazarenko et L. Audibert. Formal Description of Resources for Ontology-based Semantic Annotation. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 3765–3772, Malte, mai 2010.
 - [CI-49] T. Poibeau et C. Messiant. Automatic lexical acquisition from corpora : some limitations and tentative solutions. *Proceedings of eLexicography in the 21st century : new challenges, new applications (eLEX2009)*, pages 241–248, 2010.
 - [CI-48] L. Sun, A. Korhonen, T. Poibeau et C. Messiant. Investigating the cross-linguistic potential of VerbNet-style classification. *Proceedings of the 23rd International Conference on Computational Linguistics (CoLing 2010)*, pages 1056–1064, Beijing, China, Chine, 2010.
 - [CI-47] R. Wolfler Calvo, A. Rozenknop, D. Chemla, S., L. Alfandari, L. Létocart et G. Turri. A column generation approach for scheduling nuclear power plants refueling. *Proceedings of the 24th European conference on operational research (EURO 2010)*, Portugal, juillet 2010.
 - [CI-46] G. Xiao, Z. Lin, Y. Ma et G. Qi. Computing Inconsistency Measurements under Multi-Valued Semantics by Partial Max-SAT Solvers. *Proceedings of the 12th International Conference on the Principles of Knowledge Representation and Reasoning (KR-2010)*, pages 340–349, Canada, mai 2010.
 - [CI-45] H. Zargayouna et A. Nazarenko. Evaluation of Textual Knowledge Acquisition Tools : a Challenging Task. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 435–440, Valletta, Malte, 2010.
- 2009**
- [CI-44] A. Bossard et T. Poibeau. Integrating Document Structure into a Multi-Document Summarizer. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 09)*, pages 45–49, Borovets, Bulgarie, septembre 2009.
 - [CI-43] A. Bouffier et T. Poibeau. Analyzing the Scope of Conditions in Texts : A Discourse-Based Approach. *Proceedings of the 11th Conference of the Pacific Association for Computational Linguistics (PACLing 2009)*, pages 1–8, Sapporo, France, 2009.
 - [CI-42] S. Boutouhami et D. Kayser. Strategies to defend a protagonist of an event. *Proceedings of the 21st International Conference on Tools with Artificial Intelligence (ICTAI 2009)*, pages 469–476, Newark, États-Unis, novembre 2009.
 - [CI-41] F. Cailliau, A. Giraudel et B. Arnulphy. Tracking out-of-date newspaper articles. *Proceedings of the Conference on Intelligent text processing and Computational Linguistics (CICLing 2009)*, pages 1–12, Mexico, Mexique, mars 2009.
 - [CI-40] V. Dragos et A. Nazarenko. Towards a semantic model to enhance knowledge sharing and discovery in organic chemistry. *Proceedings of the IADIS International Conference on Information Systems (IS 2009)*, pages 1–4, Barcelona, Espagne, 2009. 4 pages.
 - [CI-39] O. Hamon, C. Fügen, D. Mostefa, V. Arranz, M. Kolss, A. Waibel et K. Choukri. End-to-end Evaluation in Simultaneous Translation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 345–353, Athens, Greece, March–April 2009.
 - [CI-38] Y. Ma et H. Pascal. Paraconsistent reasoning for OWL 2. *Proceedings of the 3rd International Conference on Web Reasoning and Rule Systems (RR 2009)*, pages 197–211, États-Unis, octobre 2009.
 - [CI-37] Y. Ma, G. Qi, G. Xiao, P. Hitzler et Z. Lin. An Anytime Algorithm for Computing Inconsistency Measurement. *Proceedings of 3rd International Conference on Knowledge Science, Engineering and management (KSEM'09)*, pages 29–40, Autriche, novembre 2009.
 - [CI-36] E. Moreau et I. Tellier. The Crotal SRL System : a Generic Tool Based on Tree-structured CRF. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009 Shared Task)*, pages 91–96, Boulder, Colorado, États-Unis, juin 2009. Association for Computational Linguistics. ANR-07-MDCO-003.

- [CI-35] A. Nazarenko et H. Zargayouna. Evaluating term extraction. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 09)*, pages 299–304, Borovets, Bulgarie, 2009.
- [CI-34] C. Recanati, N. Rogovschi et Y. Bennani. Hybrid Unsupervised Learning to Uncover Discourse Structure. Z. Vetulani et H. Uszkoreit, éditeurs, *Proceedings of the 4th Language and Technology Conference (LTC 2009)*, volume 5603/2009 de *Lecture Notes in Artificial Intelligence, Computer Science*, pages 258–269, Poznan, Pologne, août 2009. Springer. volume of the best papers of LTC'07.
- [CI-33] S. Szulman, J. Charlet, N. Aussenac-Gilles, A. Nazarenko, E. Sardet et H. Teguiak. DA-FOE : an Ontology Building Platform From Text or Thesauri. *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD 2009)*, pages 372–375, Madeira, Portugal, octobre 2009.
- [CI-32] L. Zhou, H. Huang, G. Qi, Y. Ma, Z. Huang et Y. Qu. Measuring Inconsistency in DL-Lite Ontologies. *Proceedings of the 2009 IEEE / WIC / ACM International Conferences on Web Intelligence (WI'09)*, pages 349–356, Milan, Italie, septembre 2009.
- 2008** [CI-31] S. Benferhat, J.-F. Bonnefon, S. Boutouhami, P. Chassy, R. Da Silva Neves, D. Dubois, F. Dupin De Saint-Cyr, D. Kayser, F. Nouioua, H. Prade et S. Smaoui. A Comparative Study of Six Formal Models of Causal Ascription. S. Greco et T. Lukasiewicz, éditeurs, *Proceedings of the 2nd International Conference on Scalable Uncertainty Management (SUM 2008)*, LNAI 5291, pages 47–62, Naples, Italie, octobre 2008. Springer.
- [CI-30] A. Bossard, M. Généreux et T. Poibeau. Description of the LIPN System at TAC 2008 : Summarizing Information and Opinions. *Proceedings of the 2008 Text Analysis Conference (TAC 2008)*, pages 282–291, Gaithersburg, États-Unis, novembre 2008.
- [CI-29] M. Garnier-Rizet, G. Adda, F. Cailliau, J.-L. Gauvain, S. Guillemin-Lanne, L. Lamel, S. Vanni et C. Waast-Richard. CallSurf - Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pages 2623–2628, Marrakech, Maroc, mai 2008.
- [CI-28] N. Grabar, M.-C. Jaulent et T. Hamon. Combination of endogenous clues for profiling inferred semantic relations : experiments with Gene Ontology. *Proceedings of the AMIA Annual Symposium (AMIA 2008)*, pages 252–256, États-Unis, novembre 2008.
- [CI-27] O. Hamon et D. Mostefa. An Experimental Methodology for an End-to-End Evaluation in Speech-to-Speech Translation. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pages 39–42, Marrakech, Morroco, May 2008. ELRA.
- [CI-26] O. Hamon et D. Mostefa. The Impact of Reference Quality on Automatic MT Evaluation. *Proceedings of the 22nd International Conference on Computational Linguistics (CoLing 2008) (Poster)*, pages 39–42, Manchester, UK, August 2008.
- [CI-25] T. Hamon et N. Grabar. Acquisition of elementary synonym relations from biological structured terminology. A. Gelbukh, éditeur, *Proceedings of the Conference on Intelligent text processing and Computational Linguistics (CICLing 2008)*, numéro 4919 LNCS, pages 40–51, Haïfa, Israël, février 2008. Springer-Verlag Berlin Heidelberg.
- [CI-24] T. Hamon et N. Grabar. How Can the Term Compositionality Be Useful for Acquiring Elementary Semantic Relations? B. Nordström et A. Ranta, éditeurs, *Proceedings of the Advances in Natural Language Processing (6th International Conference on NLP) (GoTAL 2008)*, numéro 5221 LNAI, pages 809–814, Gothenburg, Suède, août 2008. Springer-Verlag Berlin Heidelberg.
- [CI-23] C. Messiant, A. Korhonen et T. Poibeau. LexSchem : A Large Subcategorization Lexicon for French Verbs. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pages 142–149, Marrakech, Maroc, 2008.
- [CI-22] F. Nouioua. A heuristic approach to order events in narrative texts. *International Symposium on Temporal Representation and Reasoning*, volume 0, pages 67–71, Los Alamitos, CA, USA, 2008. IEEE Computer Society.
- [CI-21] T. Poibeau. Controversies and Misunderstandings about Meaning : On the reception of Odgen and Richards' book, « The Meaning of Meaning ». *Proceedings of the International Conference on the History of the Language Sciences (ICoHLS'2008)*, Potsdam, Allemagne, septembre 2008. Electronic publication.

- [CI-20] T. Poibeau et C. Messiant. Do we still Need Gold Standards for Evaluation? *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pages 547–552, Maroc, mai 2008.
- [CI-19] C. Recanati. Hybrid Reasoning and the Future of Iconic Representations. P. Wang, B. Goertzel et S. Franklin, éditeurs, *Proceedings of the 1st Artificial General Intelligence Conference (AGI 2008)*, volume 171 de *Frontiers in Artificial Intelligence and Applications*, pages 299–310, Memphis, Tennessee, États-Unis, février 2008. IOS Press. pp. 299-310.
- [CI-18] E. Villemonte de la Clergerie, O. Hamon, D. Mostefa, C. Ayache, P. Paroubek et A. Vilnat. PASSAGE : from French Parser Evaluation to Large Sized Treebank. ELRA, éditeur, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pages 3570–3577, Marrakech, Morocco, May 2008. ELRA.
- 2007** [CI-17] A. Bossard et T. Poibeau. Adaptation d'une ressource prédicative pour l'extraction d'information. *Proceedings of the Lexis and Grammar Conference (LGC 2007)*, pages 1–8, Bonifacio, France, 2007. LADL.
- [CI-16] A. Bouffier. From Texts to Structured Documents : The Case of Health Practice Guidelines. Springer, éditeur, *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference 2007 (ISWC/ASWC 2007)*, pages 916–920, Busan, Corée, République De, novembre 2007.
- [CI-15] A. Bouffier et T. Poibeau. Restructuring documents : A discourse based approach. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 07)*, Bulgarie, 2007.
- [CI-14] A. Delteil, L. Bourgeois et F. Lévy. A formalism for active and informative web services. *International Conference on Internet and Web applications and services, ICIW*, pages 28–, Washington, DC, USA, May 2007. IEEE Computer Society.
- [CI-13] S. Després et V. Ceașu. Semantic case-based reasoning framework for text categorization. *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference 2007 (ISWC/ASWC 2007)*, pages 736–749, Corée, République De, novembre 2007.
- [CI-12] S. Després et S. Mhiri. Ontology Usability via a Visualization Tool for the Semantic Indexing of Medical Reports (DICOM SR). *Proceedings of Usability & HCI for Medicine and Health Care3 (USAB 07)*, pages 409–414, Autriche, novembre 2007.
- [CI-11] S. Després et S. Mhiri. Ontology Visualization Tool for Indexing DICOM Structured Reporting (SR) Documents. *Proceedings of the 10th International Protégé Conference*, pages 150–153, Hongrie, juillet 2007.
- [CI-10] P. Estrella, O. Hamon et A. Popescu-Belis. How Much Data is Needed for Reliable MT Evaluation? Using Bootstrapping to Study Human and Automatic Metrics. *Proceedings of the Machine Translation Summit XI*, pages 167–174, Copenhagen, Denmark, September 2007.
- [CI-9] O. Hamon, A. Hartley, A. Popescu-Belis et K. Choukri. Assessing Human and Automated Quality Judgments in the French MT Evaluation Campaign CESTA. *Proceedings of the Machine Translation Summit XI*, pages 167–174, Copenhagen, Denmark, September 2007.
- [CI-8] T. Hamon, J. Derivière et A. Nazarenko. Ogmios : a scalable NLP platform for annotating large web document collections. *Proceedings of the Corpus Linguistics Conference (CL 2007)*, pages 195–206, Birmingham, Royaume-Uni, juillet 2007. Electronic publication. ALVIS STREP Project.
- [CI-7] T. Hamon, A. Nazarenko, T. Poibeau, S. Aubin et J. Derivière. A Robust Linguistic Platform for Efficient and Domain specific Web Content Analysis. *Proceedings of the 8th International Conference on Computer-Assisted Information Retrieval (RIA0 2007)*, pages 1–15, Pittsbergh, États-Unis, mai 2007. Edition sur CDROM. ALVIS STREP Project.
- [CI-6] T. Mondary, A. Bouffier et A. Nazarenko. Between browsing and search, a new model for navigating through large documents. S. Vosniadou, D. Kayser et A. Protopapas, éditeurs, *Proceedings of the European Cognitive Science Conference (EuroCogSci07)*, pages 634–639, Delphi, Grèce, mai 2007. Lawrence Erlbaum Associates.

- [CI-5] C. Recanati. Characteristics of diagrammatic reasoning. S. Vosniadou, D. Kayser et A. Protopapas, éditeurs, *Proceedings of the Second European Cognitive Science Conference (EuroCogSci07)*, pages 510–515, Delphi, Grèce, mai 2007. Lawrence Erlbaum Associates.
- [CI-4] C. Recanati, N. Rogovschi et Y. Bennani. Sequencing of verbs - a study on tense and aspect using unsupervised learning. G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov et N. Nikolov, éditeurs, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 07)*, pages 499–503, Borovets, Bulgarie, septembre 2007. Institute for Parallel Processing, Bulgarian Academy of Sciences.
- [CI-3] C. Recanati, N. Rogovschi et Y. Bennani. The structure of verbal sequences analyzed with unsupervised learning techniques. Z. Vetulani, éditeur, *Proceedings of the 3rd Language and Technology Conference (LTC 2007)*, pages 325–329, Poznan, Pologne, octobre 2007. Wydawnictwo Poznanskie Sp. z o.o. with co-operation of Fundacja Uniwersytetu im. A. Mickiewicza.
- [CI-2] D. Weissenbacher et A. Nazarenko. A Bayesian approach combining surface clues and linguistic knowledge : Application to the anaphora resolution problem. G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov et N. Nikolov, éditeurs, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 07)*, pages 1–7. Institute for Parallel Processing, Bulgarian Academy of Sciences, septembre 2007.
- [CI-1] D. Weissenbacher et A. Nazarenko. A bayesian classifier for the recognition of the impersonal occurrences of the 'it' pronoun. *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC'07)*, pages 145–150, Portugal, mai 2007.

Workshops et autres conférences avec comité de lecture

- 2012** [CO-99] I. Bannour et H. Zargayouna. Une plate-forme open-source de recherche d'information sémantique. *Actes de la Conférence en Recherche d'Information et Applications (CORIA 2012)*, pages 167–178, Bordeaux, France, mars 2012.
- [CO-98] S. Bannour et L. Audibert. Vers une approche interactive pour l'annotation sémantique. *Actes des 23èmes journées francophones d'Ingénierie des Connaissances (IC'12)*, Paris, France, mai 2012.
- [CO-97] S. Ben Abbès, A. Scheuermann, T. Meilender et M. D'Aquin. Characterizing Modular Ontologies. *Proceedings of the 6th International Workshop on Modular Ontologies*, pages 13–25, Graz, Autriche, juillet 2012.
- [CO-96] S. Ben Abbès, H. Zargayouna et A. Nazarenko. Une cartographie de domaine pour guider la construction d'une ressource sémantique. *Actes des 23èmes journées francophones d'Ingénierie des Connaissances (IC'12)*, pages 100–103, Paris, France, juin 2012. Quaero.
- [CO-95] C. Benzitoun, K. Fort et B. Sagot. TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. *Actes de la 19ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'12)*, pages 99–112, Grenoble, France, juin 2012.
- [CO-94] A. Bossard et C. Rodrigues. Combining a Multi-document Summarization System with a Genetic Algorithm. Springer, éditeur, *Proceedings of the 3rd International Workshop on Combinations of Intelligent Methods and Applications (CIMA 2012)*, volume 8 de *Smart Innovation, Systems and Technologies*, pages 71–87, Arras, France, France, janvier 2012. Springer.
- [CO-93] K. Fort et V. Claveau. Annotation manuelle de matchs de foot : Oh la la la ! l'accord inter-annotateurs ! et c'est le but ! *Actes de la 19ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'12)*, pages 383–390, Grenoble, France, juin 2012.
- [CO-92] J. Le Roux, B. Favre, A. Nasr et S. A. Mirroshandel. Generative Constituent Parsing and Discriminative Dependency Reranking : Experiments on English and French. *Proceedings of the ACL Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)*, pages 89–99, Corée, République De, 2012.

- [CO-91] J. Le Roux, J. Foster, J. Wagner, S. Z. Kaljahi, Rasul et A. Bryl. DCU-Paris13 Systems for the SANCL 2012 Shared Task. *Proceedings of the NAACL 2012 First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, pages 1–4, Montréal, Canada, juin 2012.
- [CO-90] J. Le Roux, B. Sagot et D. Seddah. Statistical Parsing of Spanish and Data Driven Lemmatization. *ACL Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)*, pages 55–61, Jeju, République De Corée, 2012.
- [CO-89] N. Mimouni, A. Nazarenko et S. Salotti. Analyse formelle et relationnelle de concepts pour la modélisation et l’interrogation d’une collection documentaire. *Actes de EGC 4ème Atelier Recherche d’Information SEMantique (RISE 2012)*, pages 46–57, Bordeaux, France, janvier 2012.
- [CO-88] N. Mimouni, A. Nazarenko et S. Salotti. Classification conceptuelle d’une collection documentaire. Intertextualité et Recherche d’Information. *Actes de la Conférence en Recherche d’Information et Applications (CORIA 2012)*, pages 123,134, Bordeaux, France, mars 2012.
- [CO-87] N. Omrane, A. Nazarenko et S. Szulman. Comment guider le travail de normalisation terminologique? *Actes des 23èmes journées francophones d’Ingénierie des Connaissances (IC’12) (poster)*, Paris, France, juin 2012.
- [CO-86] S. Rosset, C. Grouin, K. Fort, O. Galibert, J. Kahn et P. Zweigenbaum. Structured Named Entities in two distinct press corpora : Contemporary Broadcast News and Old Newspapers. *Proceedings of the ACL 6th Linguistic Annotation Workshop (LAW 2012)*, pages 40–48, Jeju, Corée, République De, juillet 2012.
- [CO-85] J. Van Puymbrouck, A. Nazarenko et A. Rozenknop. Construction de profils folksonomiques pour la Recherche d’Information. *Actes de la Conférence en Recherche d’Information et Applications (CORIA 2012)*, pages 71–82, Bordeaux, France, mars 2012.
- 2011** [CO-84] S. Bannour, L. Audibert et A. Nazarenko. Mesures de similarité distributionnelle entre termes. *Actes des 22èmes journées francophones d’Ingénierie des Connaissances (IC’11)*, pages 523–538, Chambéry, France, mars 2011.
- [CO-83] A. Bossard. Génération de résumés de mise à jour : Utilisation d’un algorithme de classification non supervisée pour détecter la nouveauté dans les articles de presse. *Actes d’l’atelier CIDN - Clustering Incrémental et Méthodes de Détection de Nouveauté*, Brest, France, janvier 2011.
- [CO-82] C. Grouin, S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert et L. Quintard. Proposal for an Extension of Traditional Named Entities : from Guidelines to Evaluation, an Overview. *Proceedings of the ACL 5th Linguistic Annotation Workshop (LAW 2011)*, pages 92–100, Portland, États-Unis, juin 2011.
- [CO-81] A. Guissé, F. Lévy et A. Nazarenko. Un moteur sémantique pour explorer des textes régmentaires. *Actes des 22èmes journées francophones d’Ingénierie des Connaissances (IC’11)*, pages 1–8, Chambéry, France, 2011.
- [CO-80] J. Jourde, A.-P. Manine, P. Veber, K. Fort, R. Bossy, E. Alphonse et P. Bessières. BioNLP Shared Task 2011 - Bacteria Gene Interactions and Renaming. *ACL Workshop on Biomedical Natural Language Processing (BioNLP 2011)*, pages 65–73, Portland, États-Unis, juillet 2011.
- [CO-79] D. Kayser. Ressources lexicales pour une sémantique inférentielle : un exemple, le mot « quitter ». *Actes de la 18ème conférence sur le Traitement Automatique des Langues Naturelles (TALN’11) (poster)*, Montpellier, France, 2011.
- [CO-78] Y. Ma et F. Lévy. Computing Fine-grained Semantic Annotations of Texts. *Proceedings of the 10th International Semantic Web Conference Poster Demo session (ISWC-PD’11)*, pages 1–4, Allemagne, octobre 2011.
- [CO-77] N. Omrane, A. Nazarenko, P. Rosina, S. Szulman et C. Westphal. Lexicalized ontology for the management of business rules : An industrial experiment. *Actes de la 9ème conférence « Terminologie et Intelligence Artificielle » (TIA 2011) - Workshop « Ontologies and lexicon : new insights »*, pages 9–12, Paris, France, novembre 2011.
- [CO-76] N. Omrane, A. Nazarenko et S. Szulman. Le poids des entités nommées dans le filtrage des termes d’un domaine. *Actes de la 9ème conférence « Terminologie et Intelligence Artificielle » (TIA 2011)*, pages 80–86, Paris, France, septembre 2011.

- [CO-75] N. Omrane, A. Nazarenko et S. Szulman. Les entités nommées : des clés linguistiques pour la conceptualisation. *Actes des 22èmes journées francophones d'Ingénierie des Connaissances (IC'11)*, pages 435–450, Chambéry, France, mai 2011.
- [CO-74] B. Sagot, K. Fort, G. Adda, J. Mariani et B. Lang. Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. *Actes de la 18ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'11)*, pages 199–210, Montpellier, France, 2011.
- [CO-73*] D. Seddah, J. Le Roux et B. Sagot. Towards Using Data Driven Lemmatization for Statistical Constituent Parsing of Italian. *Proceedings of the Evaluation of NLP and Speech Tools for Italian (EVALITA 2011)*, page 6 pages (electronic publication), Rome, Italie, décembre 2011.
- [CO-72] S. Szulman. Une nouvelle version de l'outil Terminae de construction de ressources termino-ontologiques. *22èmes journées francophones d'Ingénierie des Connaissances (IC'11 (poster))*, page 3 pages, Chambéry, France, mai 2011.
- [CO-71] R. Wolfler Calvo, A. Rozenknop, D. Chemla, L. Alfandari et L. Létocart. A column generation approach for scheduling nuclear power plants refueling. *Workshop following the challenge EURO/ROADEF 2010*, Clamart, France, mars 2011.
- 2010** [CO-70] S. Ben Abbès, H. Zargayouna et A. Nazarenko. Évaluation de classes sémantiques pour la construction d'ontologies. S. Després, éditeur, *Actes des 21èmes journées francophones d'Ingénierie des Connaissances (IC'10)*, pages 297–308, Nîmes, France, 2010. Ecole des Mines d'Alès.
- [CO-69] R. Bentebibel, A. Nazarenko et S. Szulman. Un système d'aide à l'extraction de relations sémantiques pour la construction d'ontologies à partir de textes. *Actes des 10èmes journées francophones « Extraction et Gestion des Connaissances » (EGC 2010)*, volume RNTI-E-19, pages 483–494, Hammamet, France, janvier 2010.
- [CO-68] S. Boutouhami et D. Kayser. Comment construire une description favorisant un protagoniste d'un événement? *Reconnaissance des Formes et Intelligence Artificielle (RFIA'10)*, pages 670–677, Caen, France, janvier 2010.
- [CO-67] K. Fort, C. François et M. Ghribi. Evaluer des annotations manuelles dispersées : les coefficients sont-ils suffisants pour estimer l'accord inter-annotateurs? *Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, France, 2010.
- [CO-66] K. Fort et B. Sagot. Influence of Pre-annotation on POS-tagged Corpus Development. *Proceedings of the ACL 4th Linguistic Annotation Workshop (LAW 2010)*, pages 56–63, Uppsala, Suède, 2010.
- [CO-65] Y. Ma et P. Hitzler. Distance-based Measures of Inconsistency and Incoherency for Description Logics. *Proceedings of the 23rd International Workshop on Description Logics (DL'10)*, pages 475–485, Canada, mai 2010.
- [CO-64] N. Omrane, A. Nazarenko et S. Szulman. Combining terms and named entities for modeling domain ontologies from texts. *Proceedings of the 17th International conference on Knowledge Engineering and Knowledge Management (EKAW Demo 2010)*, pages 3–4, Lisbon, Portugal, octobre 2010.
- [CO-63] N. Omrane, A. Nazarenko et S. Szulman. Les entités nommées : éléments pour la conceptualisation. *Actes des 21èmes journées francophones d'Ingénierie des Connaissances (IC'10)*, Nîmes, France, juin 2010.
- [CO-62] S. Szulman, N. Aussenac-Gilles, J. Charlet, A. Nazarenko, E. Sardet et H. Teguiak. DA-FOE : A Platform for Building Ontologies from Texts. *The 17th International conference on Knowledge Engineering and Knowledge Management (EKAW Demo 2010)*, pages 95–96, Lisbon, Portugal, octobre 2010.
- 2009** [CO-61] R. Bentebibel, A. Nazarenko et S. Szulman. Mise en lumière de relations sémantiques pour la construction d'ontologies à partir de textes. *8ème conférence « Terminologie et Intelligence Artificielle » (TIA 2009) (poster)*, Toulouse, France, novembre 2009.
- [CO-60] A. Bossard. Une Approche Mixte -statistique et structurelle- pour le Résumé Automatique. *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, pages 1–10, Senlis, France, juin 2009.

- [CO-59] A. Bossard, M. Généreux et T. Poibeau. CBSEAS, a Summarization System - Integration of Opinion Mining Techniques to Summarize Blogs. *Proceedings of the Demonstrations Session at the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 5–8, Athènes, Grèce, mars 2009.
- [CO-58] S. Boutouhami. Techniques argumentatives pour aider à générer des descriptions orientées d'un événement. *Actes de la 13ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'09)*, Senlis, France, 2009.
- [CO-57] S. Boutouhami et D. Kayser. A system for generating biased descriptions. *Proceedings of Extended abstracts of the International Multidisciplinary Workshop on Causality*, pages 69–72, Toulouse, France, juin 2009.
- [CO-56] S. Després et S. Szulman. Choisir une ontologie noyau en vue de sa réutilisation Une stratégie d'évaluation. L. Bellatrèche, G. Kassel et P. Thiran, éditeurs, *Actes des 3èmes Journées Francophones sur les Ontologies (JFO 2009)*, pages 131–139, Poitiers, France, décembre 2009.
- [CO-55] S. Després et H. Zargayouna. Evaluation of knowledge based applications : benchmark and guidelines. *Proceedings of the 2d international workshop on Knowledge Acquisition, Reuse and Evaluation (KARE 2009)*, pages 472–478, Marrakech, Maroc, 2009.
- [CO-54] S. Després, H. Zargayouna et R. Bentebibel. Quelles connaissances pour se mettre à TAAABLE? *Actes du 17ème atelier sur le Raisonnement à Partir de Cas (RàPC'09)*, pages 151–168, Paris, France, 2009.
- [CO-53] K. Fort, M. Ehrmann et A. Nazarenko. Towards a Methodology for Named Entities Annotation. *Proceedings of the ACL 3rd Linguistic Annotation Workshop (LAW 2009)*, pages 142–145, Singapour, 2009. Quaero.
- [CO-52] K. Fort, M. Ehrmann et A. Nazarenko. Vers une méthodologie d'annotation des entités nommées en corpus? *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France, 2009. Quaero.
- [CO-51] A. Guisé, F. Lévy, A. Nazarenko et S. Szulman. Annotation sémantique de documents pour la révision des règles métiers. *Actes de la 8ème conférence « Terminologie et Intelligence Artificielle » (TIA 2009)*, Toulouse, France, novembre 2009.
- [CO-50] F. Hamdi, B. Safar, H. Zargayouna et C. Reynaud. Partitionnement d'ontologies pour le passage à l'échelle des techniques d'alignement. *Actes des 9èmes journées francophones « Extraction et Gestion des Connaissances » (EGC 2009)*, pages 409–420, Strasbourg, France, 2009. Cepadues.
- [CO-49] D. Kayser et F. Lévy. Causality : Purposes, Core Notions, Properties. *Extended abstracts of the International Multidisciplinary Workshop on Causality*, pages 24–27, Toulouse, France, juin 2009.
- [CO-48] Y. Ma et L. Audibert. Détection des contradictions dans les annotations sémantiques. *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, pages 609–618, Senlis, France, juin 2009.
- [CO-47] Y. Ma, L. Audibert et A. Nazarenko. Ontologies étendues pour l'annotation sémantique. *Actes des 20èmes journées francophones d'Ingénierie des Connaissances (IC'09)*, pages 205–216, Hammamet, Tunisie, Tunisie, mai 2009. Quaero, financé par OSEO, agence nationale de valorisation de la recherche.
- [CO-46] C. Messiant, T. Nakamura et S. Voyatzi. La complémentarité des approches manuelle et automatique en acquisition lexicale. *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, pages 1–10, Senlis, France, 2009.
- [CO-45] C. Messiant et T. Poibeau. Acquisition d'informations lexicales à partir de corpus. *Actes du 3ème colloque international de l'Association Française de Linguistique Cognitive (AFLICO)*, pages 1–2, Nanterre, France, mai 2009. 3 pages.
- [CO-44] T. Mondary et S. Després. Analyse de concepts formels pour la construction d'ontologies à partir de textes : la question du corpus. N. Laliche, A. Braud, P. Gançarski et J.-G. Ganascia, éditeurs, *Actes de l'Atelier « Qualité des Données et des Connaissances » des 9èmes journées francophones Extraction et Gestion des connaissances (EGC09, QDC 2009)*, pages 1–8, Strasbourg, France, janvier 2009.

- [CO-43] E. Moreau, I. Tellier, A. Balvet, G. Laurence, A. Rozenknop et T. Poibeau. Annotation fonctionnelle de corpus arborés avec des Champs Aléatoires Conditionnels. *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, pages 1–10, Senlis, France, juin 2009. Crotal - ANR-07-MDCO-03.
- [CO-42] F. Nouioua et D. Kayser. A Norm-Based Approach to Causal Reasoning. *Extended abstracts of the International Multidisciplinary Workshop on Causality*, pages 62–65, Toulouse, France, juin 2009.
- [CO-41] H. Zargayouna et A. Nazarenko. Peut-on évaluer les outils d'acquisition de connaissances à partir de textes? *Actes de l'atelier sur l'Evaluation des méthodes d'Extraction de Connaissances dans les Données (EvalECD'09)*, pages 5–16, Strasbourg, France, 2009.
- [CO-40] X. Zhang, G. Qi, Y. Ma et Z. Lin. Quasi-Classical Semantics for Expressive Description Logics. *Proceedings of the 22nd International Workshop on Description Logics (DL'09)*, page 11, Royaume-Uni, juillet 2009.
- 2008**
- [CO-39] F. Badra, R. Bendaoud, R. Bentebibel, P.-A. Champin, J. Cojan, A. Cordier, S. Després, S. Jean-Daubias, J. Lieber, T. Meilender, A. Mille, E. Nauer, A. Napoli et Y. Toussaint. TAAABLE : Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. M. Schaaf, éditeur, *9th European Conference on Case-Based Reasoning - ECCBR 2008, Workshop Proceedings*, pages 219–228, Trier, Allemagne, 2008.
- [CO-38] R. Besançon, S. Chaudiron, D. Mostefa, O. Hamon, I. Timimi et K. Choukri. Overview of CLEF 2008 INFILE pilot track. *Proceedings of the CLEF 2008 Working Notes*, pages 1–7, Aarhus, Denmark, September 2008.
- [CO-37] A. Bossard et T. Poibeau. Regroupement Automatique de Documents en Classes Événementielles. *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'08)*, pages 201–210, Avignon, France, juin 2008.
- [CO-36] A. Bouffier, C. Duclos et T. Poibeau. Analyse et structuration automatique des guides de bonnes pratiques cliniques : essai d'évaluation. *Actes des 19èmes journées francophones d'Ingénierie des Connaissances (IC'08)*, pages 135–146, Nancy, France, juin 2008.
- [CO-35] F. Cailliau et A. Giraudel. Enhanced search and navigation on conversational speech. *SIGIR Workshop : Searching Spontaneous Conversational Speech (SSCS 2008)*, pages 66–70, Singapour, 2008.
- [CO-34] F. Cailliau et C. Poudat. Caractérisation lexicale des contributions clients agents dans un corpus de conversations téléphoniques retranscrites. *9èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008)*, pages 267–276, Lyon, France, mars 2008.
- [CO-33] F. Cailliau, C. Poudat et A. Giraudel. Fusion de ressources hétérogènes pour la recherche d'information. *Conférence en Recherche d'Information et Applications (CORIA 2008)*, pages 433–440, Trégastel, France, mars 2008.
- [CO-32] J. Charlet, S. Szulman, G. Pierra, N. Nadah, H. Teguiak, Valery, N. Aussenac-Gilles et A. Nazarenko. DAFOE : A Multimodel and Multimethod Platform for Building Domain Ontologies. *Actes des 2èmes Journées Francophones sur les Ontologies (JFO 2008)*, page 66, Lyon, France, novembre 2008. 12 pages.
- [CO-31] S. Després. Construction d'ontologies à partir de textes dans le domaine juridique. *Actes du colloque annuel de la Société Française de Terminologie (SFT)*, France, 2008.
- [CO-30] S. Després et S. Szulman. Réseau terminologique versus Ontologie. *Actes de Terminologie & Ontologie : Théories et applications (TOTh 2008)*, pages 17–34, France, juillet 2008.
- [CO-29] S. Després et S. Szulman. Sémantique et réutilisation d'ontologie générique. *Actes des 8èmes journées francophones « Extraction et Gestion des Connaissances » (EGC 2008)*, pages 121–126, Sophia-Antipolis, France, 2008.
- [CO-28] M. Généreux, T. Poibeau et M. Koppel. Sentiment analysis using automatically labelled financial news. *Proceedings of the LREC Workshop on Sentiment Analysis : Emotion, Metaphor, Ontology and Terminology (EMOT 2008)*, Marakech, Maroc, juin 2008.
- [CO-27] N. Grabar, T. Hamon et T. Dart. Term variation and semantics for document classification and detection of obesity and its co-morbidities cases. *Proceedings of the AMIA Second i2b2 Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data*, pages 252–256, États-Unis, novembre 2008.

- [CO-26] N. Grabar, P.-C. Varoutas, P. Rizand, A. Livartowski et T. Hamon. Automatic acquisition of synonyms from French UMLS for enhanced search of EHRs. *Proceedings of the 21st International Congress of the European Federation for Medical Informatics (MIE 2008)*, volume 136 de *Studies in health technology and informatics*, pages 809–14, Göteborg, Suède, mai 2008. IOS Press.
- [CO-25] F. Hamdi, H. Zargayouna, B. Safar et C. Reynaud. TaxoMap in the OAEI 2008 alignment contest. *Proceedings of the 3rd International Workshop on Ontology Matching (OM'08)*, pages 206–213, Karlsruhe, Allemagne, 2008.
- [CO-24] O. Hamon, D. Mostefa et V. Arranz. Diagnosing human judgments in MT evaluation : an example based on the Spanish language. *Proceedings of Mixing Approaches to Machine Translation (MATMT 2008)*, pages 19–26, San Sebastian, Spain, February 2008.
- [CO-23] N. Kamenoff. Processus de génération d'une animation en 3 dimensions à partir d'un texte. *19èmes journées francophones d'Ingénierie des Connaissances (IC'08)*, pages 147–159, Nancy, France, juin 2008.
- [CO-22] C. Messiant. ASSCI : A Subcategorization Frames Acquisition System for French Verbs. *Proceedings of the Association for Computational Linguistics (ACL, Student Research Workshop)*, pages 55–60, Columbus, Ohio, Maroc, 2008.
- [CO-21] T. Mondary, S. Després, A. Nazarenko et S. Szulman. Construction d'ontologies à partir de textes : la phase de conceptualisation. *Proceedings of the 19èmes journées francophones d'Ingénierie des Connaissances (IC'08)*, pages 87–98, Nancy, France, juin 2008.
- [CO-20] T. Poibeau. Quelques remarques sur le « champ linguistique » en France dans l'après guerre (1950-1970). *Actes du Congrès Mondial de linguistique française (CMLF 2008)*, Paris, France, juillet 2008.
- [CO-19] A. Vilnat, G. Francopoulo, O. Hamon, S. Loiseau, P. Paroubek et E. Villemonte de la Clergerie. Large Scale Production of Syntactic Annotations to Move Forward. *Proceedings of the CoLing Workshop on Cross-Framework and Cross-Domain Parser Evaluation (CoLing 2008)*, pages 36–43, Manchester, UK, August 2008.
- 2007** [CO-18] L. Audibert. Désambiguïsation lexicale automatique : sélection automatique d'indices. N. Hathout et P. Muller, éditeurs, *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'07)*, pages 13–22, Toulouse, France, juin 2007. IRIT Press.
- [CO-17] A. Bossard. Vers une Ressource Prédicative pour l'Extraction d'Information. *Actes de la 11ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'07)*, pages 347–356, Toulouse, France, juin 2007. IRIT Press.
- [CO-16] A. Bouffier et T. Poibeau. Automatically Restructuring Practice Guidelines using the GEM DTD. K. Bretonnel Cohen, D. Demner-Fushman, C. Friedman, L. Hirschman et J. Pestian, éditeurs, *Proceedings of the ACL Workshop on Biomedical Natural Language Processing (BioNLP 2007)*, pages 113–120, Prague, Tchèque, République, 2007. ACL.
- [CO-15] F. Cailliau et C. de Lopy. Aides à la navigation dans un corpus de transcriptions d'oral. *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'07)*, pages 143–152, Toulouse, France, juin 2007. IRIT Press.
- [CO-14] V. Ceausu et S. Després. Learning Term to Concept Mapping Through Verbs : A Case Study. *Proceedings of Semantic Authoring, Annotation and Knowledge Markup Workshop of Conference on Knowledge Capture (KCap07)*, pages 1–8, Canada, octobre 2007.
- [CO-13] S. Després, F. Fürst et S. Szulman. Construction d'une ontologie du domaine HSE. *Actes des 18èmes journées francophones d'Ingénierie des Connaissances (IC'07)*, pages 133–144, Grenoble, France, 2007.
- [CO-12] S. Després, S. Mhiri et E. Zagrouba. Unification d'ontologies hétérogènes pour l'indexation sémantique de compte rendu standardisé d'examen du patient. *Actes de l'atelier Ontologies et gestion de l'hétérogénéité sémantique (OGHS)*, pages 60–62, Grenoble, France, juillet 2007.
- [CO-11] S. Després et S. Szulman. Exploitation linguistique d'une ontologie générique. *Actes de la 7ème conférence « Terminologie et Intelligence Artificielle » (TIA 2007)*, pages 127–136, Sophia-Antipolis, France, octobre 2007.

- [CO-10] T. Hamon, J. Derivière et A. Nazarenko. Ogmios : une plate-forme d'annotation linguistique. *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'07)*, pages 103–112, Toulouse, France, juin 2007. IRIT Press. ALVIS STREP Project.
- [CO-9] L. Lamontagne, R. Bentebibel, E. Miry et S. Després. Finding Lexical Relations for the Reuse of Investigation Reports, Beyond Retrieval. *Proceedings of Fourth Workshop on Textual CBR, ICCBR'07*, pages 252–256, Irlande, août 2007.
- [CO-8] T. Mondary. Construction d'Ontologies à partir de Textes. Application au Droit International du Travail. *Session de Démonstrations et de Posters associée à la 7ème conférence Terminologie et Intelligence Artificielle (TIA'07)*, pages 7–9, Sophia-Antipolis, France, octobre 2007.
- [CO-7] T. Poibeau. Knowledge poor methods (sometimes) perform poorly. *Proceedings of the ACL Fourth International Workshop on Semantic Evaluations (SemEval 2007)*, pages 418–421, Prague, Tchèque, République, 2007. Association for Computational Linguistics.
- [CO-6] C. Recanati et N. Rogovschi. Enchaînements verbaux - étude sur le temps et l'aspect utilisant des techniques d'apprentissage non supervisé. *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'07)*, pages 379–388, Toulouse, France, juin 2007. IRIT Press. 10 pages.
- [CO-5] N. Rogovschi, Y. Bennani et C. Recanati. Apprentissage neuro-markovien pour la classification non supervisée de données structurées en séquences. *Actes de l'atelier Fouille de données et algorithmes bio-mimétiques des 7èmes journées francophones « Extraction et Gestion des Connaissances » (EGC 2007)*, pages 49–60, Namur, Belgique, janvier 2007. Cépaduès Editions.
- [CO-4] D. Weissenbacher et A. Nazarenko. Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'07)*, pages 145–155, France, juin 2007. ATALA.
- [CO-3] H. Zargayouna, O. Hamon et A. Nazarenko. Evaluation des outils terminologiques : état des lieux et propositions. *Actes de la 7ème conférence « Terminologie et Intelligence Artificielle » (TIA 2007)*, pages 1–11, Sophia Antipolis, France, octobre 2007.
- [CO-2] H. Zargayouna, V. Rosas et S. Salotti. Shallow parsing of INEX Queries. *Comparative Evaluation of XML Information Retrieval Systems - Proceedings of the 5th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2006)*, pages 284–293, Dagstuhl, Allemagne, février 2007. LNCS, Volume Number 4518.
- [CO-1] F. Nouioua. Heuristiques pour la résolution d'anaphores dans les textes d'accidents de la route. *Journée d'étude de l'Association pour le Traitement Automatique des Langues sur la « Résolution des anaphores en traitement automatique des langues » (ATALA'07)*, pages 1–4.

Direction d'ouvrages (DO)

Livres

- 2008** [LI-1] T. Poibeau et A. Condamines. *Extraction d'information : l'apport de la linguistique*. Revue Française de Linguistique Appliquée, 2008.

Édition d'ouvrages collectifs

- 2012** [ED-5*] P. Livet, D. Kayser, F. Varenne et J. Ferber. Les perspectives épistémologiques ouvertes par l'intelligence artificielle. Henri Prade et al., éditeur, *L'IA aujourd'hui*. Cepadues, 2012.
- 2009** [ED-4] A. Alishahi, T. Poibeau et A. Villavicencio. *Proceedings of the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition*. Association for Computational Linguistics, 2009.
- [ED-3] A. Nazarenko, T. Poibeau et L. Audibert. *Actes de la conférence Traitement Automatique des Langues Naturelles*. ATALA, 2009.
- 2008** [ED-2] T. Poibeau, S. Bandyopadhyay, H. Saggion et R. Yangarber. *Proceedings of the 2nd workshop on « Multi-source, Multilingual Information Extraction and Summarization »*. COLING-ACL, 2008.
- 2007** [ED-1] T. Poibeau et H. Saggion. *Proceedings of the 1st workshop on « Multi-source, Multilingual Information Extraction and Summarization »*. Recent Advances in Natural Language Processing, 2007.

Logiciels et Brevets (LO)

- 2012** [LO-5] T. Mondary. ValiTerms (software), juin 2012. ValiTerms is an HTML5 interface generator that helps the validation of terms in a corpus. ValiTerms finds their occurrences and presents them to users into a pretty html page. This software is released on SourceForge (<http://sourceforge.net/projects/valiterms>).
- [LO-4] S. Szulman. Logiciel Terminae - Version 2012, 2012. plateforme d'aide à la construction de ressources termino-ontologiques à partir de ressources textuelles. (http://lipn.fr/terminae/index.php/Main_Page).
- 2011** [LO-3] T. Mondary et J. Van Puymbrouck. Termometer (software), avril 2011. Termometer is a tool that computes terminological precision and recall between a gold standard and a list of terms, using a gradual relevance measure. This software is released on Sourceforge (<http://sourceforge.net/projects/termometerxd>).
- [LO-2] E. Moreau. LIPN UIMA Platform 0.1.* User/developer guide, mars 2011. This is a "work in progress" version funded by the Quaero Program.
- 2007** [LO-1] J.-V. Loddo et T. Hamon. Marionnet/Iutoppix : un environnement pédagogique pour la simulation de réseaux locaux, juillet 2007. Démonstration à la conférence EIAH 2007 (Environnements Informatiques pour l'Apprentissage Humain) Projet E-Learning Université Paris 13.

Thèses et Habilitations (TH)

- 2011** [TH-11] T. Mondary. Construction d'ontologies à partir de textes. L'apport de l'analyse de concepts formels. Thèse, Université Paris 13, mai 2011.
- 2010** [TH-10] A. Bossard. Contribution au résumé automatique multi-documents. Thèse, Université Paris 13, juillet 2010.
- [TH-9] S. Boutouhami. Un système de générations de descriptions argumentées. Thèse, Université Paris 13, 2010.
- [TH-8] F. Cailliau. Des ressources aux traitements linguistiques : le rôle d'une architecture linguistique. Thèse, Université Paris 13, décembre 2010.
- [TH-7] O. Hamon. Vers une architecture générique et pérenne pour l'évaluation en traitement automatique des langues : spécifications, méthodologies et mesures. Thèse, Université Paris 13, décembre 2010.
- [TH-6] C. Messiant. Acquisition automatique de schémas de sous-catégorisation à partir de corpus bruts. Thèse, Université Paris 13, novembre 2010.
- 2008** [TH-5] A. Bouffier. Analyse discursive automatique de textes, application à la modélisation de connaissances. Thèse, Université Paris 13, 2008.
- [TH-4] T. Poibeu. Des mots aux textes. Analyse sémantique pour l'accès à l'information. HDR, Université Paris 13, novembre 2008.
- [TH-3] D. Weissenbacher. Influence des annotations imparfaites sur les systèmes de Traitement Automatique des Langues, un cadre applicatif : la résolution de l'anaphore pronominale. Thèse, Université Paris 13, novembre 2008.
- 2007** [TH-2] L. Bourgeois. Représentation et comparaison de web services complexes avec des logiques dynamiques. Thèse, Université Paris 13, juin 2007.
- [TH-1] F. Nouioua. Extraction et utilisation des normes pour un raisonnement causal dans un corpus textuel. Thèse, Université Paris 13, Avril 2007.

Divers (Div)

- 2011** [Ra-6] F. Poulard, E. Moreau et L. Audibert. Vers des outils robustes et interopérables pour le TAL : la piste UIMA. avril 2011.
- 2009** [Ra-5] A. Bossard et M. Généreux. Résumé automatique de textes d'opinion. avril 2009.
- [Ra-4] S. Szulman, N. Aussenac-Gilles, A. Nazarenko et J. Charlet. Projet DaFOE4App D DIFFERENTIAL AND FORMAL ONTOLOGIES EDITOR FOR APPLICATIONS Dossier A.1 / Document A.1.1 Cahier des charges scientifique et technique de la plateforme DaFOE Spécifications, septembre 2009. Dossier A.1 / Document A.1.1 Spécifications.
- [Ra-3] S. Szulman, N. Aussenac-Gilles, A. Nazarenko et J. Charlet. Projet DaFOE4App DIFFERENTIAL AND FORMAL ONTOLOGIES EDITOR FOR APPLICATIONS. Cahier des charges scientifique et technique de la plateforme DaFOE. Chapitre Modèle de données, mai 2009. Dossier A.1 / Document A.1.1.
- 2007** [Ra-2] A. Nazarenko, S. Aubin, R. Bossy, T. Erjavec, T. Hamon, D. Mladovic, C. Nédellec, T. Poibeau, D. Weissenbacher et Q. Zhou. Final report on NLP analysis and normalization, février 2007. This document describes the ALVIS NLP line that has been developed.
- [Ra-1] M. Zimina. Typologies de textes pour le traitement automatique. Rapport Technique, Laboratoire d'Informatique de l'Université Paris Nord, 2007.

Chapitre 7

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.

Sommaire

Articles dans des revues avec comité de lecture et chapitres de livres (ACL)	76
Communications avec actes (ACT)	78
Thèses et Habilitations (TH)	81
Divers (Div)	82

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

Articles dans des revues internationales avec comité de lecture sélectif

- 2012 [↑RI-12] A. Reyes, P. Rosso et D. Buscaldi. From humor recognition to irony detection : The figurative language of social media. *Data Knowledge Engineering*, 74 :1–12, 2012.
- 2011 [↑RI-11] D. Buscaldi. Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2) :16–19, 2011.
- [↑RI-10] M. d’Aquin et A. Gangemi. Is there beauty in ontologies? *Applied Ontology*, 6(3) :165–175, 2011.
- [↑RI-9] P. Rosso, S. Correa et D. Buscaldi. Passage retrieval in legal texts. *Journal of Logic and Algebraic Programming*, 80(35) :139–153, 2011.
- 2010 [↑RI-8] A. D’Andrea, G. Ferrandino et A. Gangemi. Shared iconographical representations with ontological models. *Intelligenza Artificiale*, 4(1) :50–55, 2010.
- [↑RI-7] A. Gangemi et V. Presutti. Towards a pattern science for the semantic web. *Semantic Web*, 1(1-2) :61–68, 2010.
- 2009 [↑RI-6] D. Buscaldi, P. Rosso, J. M. Gómez et E. Sanchis. Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems (JIIS)*, 34(2) :113–134, 2009.
- 2008 [↑RI-5] D. Buscaldi et P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3) :301–313, 2008.
- [↑RI-4] A. Gangemi. Norms and plans as unification criteria for social collectives. *Autonomous Agents and Multi-Agent Systems*, 17(1) :70–112, 2008.
- [↑RI-3] V. Presutti et A. Gangemi. Identity of resources and entities on the web. *International Journal of Semantic Web and Information Systems*, 4(2) :49–72, 2008.
- 2007 [↑RI-2] J. Lehmann et A. Gangemi. An ontology of physical causation as a basis for assessing causation in fact and attributing legal responsibility. *Artificial Intelligence and Law*, 15(3) :301–321, 2007.
- [↑RI-1] J. Le Roux et G. Perrier. Modélisation de la coordination dans les grammaires d’interaction. *Traitement Automatique des Langues (TAL)*, pages 89–113, 2007.

Chapitres de livre

- 2012 [↑CL-11] A. Adamou, R. Palma, P. Haase, E. Montiel-Ponsoda, G. Aguado de Cea, A. Gomez-Perez, W. Peters et A. Gangemi. *The NeOn Ontology Models*, M. Suárez-Figueroa, A. Gómez-Pérez, E. Motta et A. Gangemi, éditeurs, *Ontology Engineering in a Networked World*. Springer, 2012.
- [↑CL-10] C. Caracciolo, J. Heguiabehere, A. Gangemi, C. Baldassarre, J. Keizer et M. Taconet. *Knowledge Management at FAO : A Case Study on Network of Ontologies in Fisheries*, M. Suárez-Figueroa, A. Gómez-Pérez, E. Motta et A. Gangemi, éditeurs, *Ontology Engineering in a Networked World*. Springer, 2012.
- [↑CL-9] V. Presutti, E. Blomqvist, E. Daga et A. Gangemi. *Pattern-Based Ontology Design*, M. Suárez-Figueroa, A. Gómez-Pérez, E. Motta et A. Gangemi, éditeurs, *Ontology Engineering in a Networked World*. Springer, 2012.
- 2011 [↑CL-8] A. Gangemi, V. Presutti et E. Blomqvist. *The Computational Ontology Perspective : Design Patterns for Web Ontologies*, G. Sartor, P. Casanovas, M. Biasiotti et M. Fernandez-Barrera, éditeurs, *Approaches to Legal Ontologies*. Springer, Berlin, 2011.
- 2010 [↑CL-7] A. Gangemi. *A Semiotic Model for Bridging Lexical and Formal Semantics*, S. Bassi, éditeur, *Atti dei Seminari Signum 2006*. Edizioni della Normale, Pisa, 2010.
- [↑CL-6] A. Gangemi. *What’s in a schema? A formal metamodel for ECG and FrameNet*, C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari et L. Prévot, éditeurs, *ONTOLEX. Studies in Natural Language Processing*. Cambridge University Press, April 2010, ISBN 978-0-521-88659-8.

- [↑CL-5] A. Gangemi, N. Guarino, C. Masolo et A. Oltramari. *Interfacing WordNet with DOLCE : towards OntoWordNet*, C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari et L. Prévot, éditeurs, *ONTOLEX*. Studies in Natural Language Processing. Cambridge University Press, April 2010, ISBN 978-0-521-88659-8.
- [↑CL-4] A. Oltramari, A. Gangemi, C.-R. Huang, N. Calzolari, A. Lenci et L. Prévot. *Synergizing Ontologies and the Lexicon : a Roadmap*, C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari et L. Prévot, éditeurs, *ONTOLEX*. Studies in Natural Language Processing. Cambridge University Press, April 2010, ISBN 978-0-521-88659-8.
- [↑CL-3] C.-R. H. L. Prévot, N. Calzolari, A. Gangemi, A. Lenci et A. Oltramari. *Ontology and the lexicon : a multi-disciplinary perspective (introduction)*, C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari et L. Prévot, éditeurs, *ONTOLEX*. Studies in Natural Language Processing. Cambridge University Press, April 2010, ISBN 978-0-521-88659-8.
- 2009** [↑CL-2] A. Gangemi et V. Presutti. *Ontology Design Patterns*, S. Staab et R. Studer, éditeurs, *Handbook on Ontologies*. Springer, 2nd édition, 2009.
- 2007** [↑CL-1] A. Gangemi et V. Presutti. *Ontology design for interaction in a reasonable enterprise*. *Handbook of Ontologies for Business Interaction*, 2007.

Communications avec actes (ACT)

Communications dans des conférences internationales avec comité de lecture

- 2012** [↑CI-21] A. Gangemi, A. G. Nuzzolese, V. Presutti, F. Draicchio, A. Musetti et P. Ciancarini. Automatic Typing of DBpedia Entities. *Proceedings of the 11th International Semantic Web Conference (ISWC'12)*, Lecture Notes in Computer Science. Springer, 2012.
- [↑CI-20] V. Presutti, F. Draicchio et A. Gangemi. Knowledge Extraction based on Discourse Representation Theory and Linguistic Frames. *Proceedings of the 18th International conference on Knowledge Engineering and Knowledge Management (EKAW'12)*, Lecture Notes in Computer Science. Springer, 2012.
- 2011** [↑CI-19] G. Cleuziou, D. Buscaldi, V. Levorato et G. Dias. A pretopological framework for the automatic construction of lexical-semantic structures from texts. *CIKM '11 : Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2453–2456, New York, NY, USA, 2011. ACM.
- [↑CI-18] A. G. Nuzzolese, A. Gangemi et V. Presutti. Gathering lexical linked data and knowledge patterns from framenet. *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP 2011)*, pages 41–48, 2011.
- [↑CI-17] A. G. Nuzzolese, A. Gangemi, V. Presutti et P. Ciancarini. Encyclopedic knowledge patterns from wikipedia links. *Proceedings of the 10th International Semantic Web Conference (ISWC'11)*, pages 520–536, 2011.
- [↑CI-16] S. Peroni, A. Gangemi et F. Vitali. Dealing with markup semantics. *Proceedings of the 7th International Conference on Semantic Systems (I-SEMANTICS 2011)*, pages 111–118, 2011.
- [↑CI-15] V. Presutti, L. Aroyo, A. Gangemi, A. Adamou, B. A. C. Schopman et G. Schreiber. A knowledge pattern-based method for linked data analysis. *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP 2011)*, pages 173–174, 2011.
- 2010** [↑CI-14] A. Adamou, V. Presutti et A. Gangemi. Kali-ma : A semantic guide to browsing and accessing functionalities in plugin-based tools. *Proceedings of the 17th International conference on Knowledge Engineering and Knowledge Management (EKAW'10)*, volume 6317 de *Lecture Notes in Computer Science*, pages 483–492. Springer, 2010.
- [↑CI-13] C. Baldassarre, E. Daga, A. Gangemi, A. M. Gliozzo, A. Salvati et G. Troiani. Semantic scout : Making sense of organizational knowledge. *Proceedings of the 17th International conference on Knowledge Engineering and Knowledge Management (EKAW'10)*, volume 6317 de *Lecture Notes in Computer Science*, pages 272–286. Springer, 2010.
- [↑CI-12] E. Blomqvist, V. Presutti, E. Daga et A. Gangemi. Experimenting with extreme design. *Proceedings of the 17th International conference on Knowledge Engineering and Knowledge Management (EKAW'10)*, volume 6317 de *Lecture Notes in Computer Science*, pages 120–134. Springer, 2010.
- [↑CI-11] C. Giuliano, A. M. Gliozzo, A. Gangemi et K. Tymoshenko. Acquiring thesauri from wikis by exploiting domain models and lexical substitution. *Proceedings of the 7th Extended Semantic Web Conference (ESWC'10)*, volume 6089 de *Lecture Notes in Computer Science*, pages 121–135. Springer, 2010.
- 2009** [↑CI-10] E. Blomqvist, A. Gangemi et V. Presutti. Experiments on pattern-based ontology design. *Proceedings of the 5th International Conference on Knowledge Capture (K-CAP 2009)*, pages 41–48. ACM, 2009.
- [↑CI-9] D. Buscaldi. Toponym ambiguity in geographical information retrieval. *SIGIR '09 : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 847–847, New York, NY, USA, 2009. ACM.
- [↑CI-8] B. Coppola, A. Gangemi, A. M. Gliozzo, D. Picca et V. Presutti. Frame detection over the semantic web. *Proceedings of the 6th European Semantic Web Conference (ESWC'09)*, volume 5554 de *Lecture Notes in Computer Science*, pages 126–142. Springer, 2009.
- [↑CI-7] A. Gangemi. Introducing Pattern-based Design for Legal Ontologies. *Proceedings of the 2009 conference on Law, Ontologies and the Semantic Web : Channelling the Legal Information Flood*, volume 188 de *Frontiers in Artificial Intelligence and Applications*, pages 53–71. IOS Press, 2009.

- [↑CI-6] J. Le Roux. Deductive Parsing in Interaction Grammars. *Proceedings of the 11th International Workshop on Parsing Technologies (IWPT 2009)*, pages 65–68, France, 2009.
- [↑CI-5] A. Reyes, D. Buscaldi et P. Rosso. An analysis of the impact of ambiguity on automatic humour recognition. *TSD '09 : Proceedings of the 12th International Conference on Text, Speech and Dialogue*, pages 162–169, Berlin, Heidelberg, 2009. Springer-Verlag.
- 2008** [↑CI-4] D. Buscaldi et P. Rosso. Geo-wordnet : Automatic georeferencing of wordnet. Nicoletta Calzolari et al., éditeur, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>.
- [↑CI-3] M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric et I. Rojas. Unsupervised learning of semantic relations for molecular biology ontologies. *Proceedings of the 2008 conference on Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, pages 91–104, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.
- [↑CI-2] V. Presutti et A. Gangemi. Content ontology design patterns as practical building blocks for web ontologies. *Proceedings of the 27th International Conference on Conceptual Modeling (ER 2008)*, volume 5231 de *Lecture Notes in Computer Science*, pages 128–141. Springer, 2008.
- 2007** [↑CI-1] A. M. Gliozzo, A. Gangemi, V. Presutti, E. Cardillo, E. Daga, A. Salvati et G. Troiani. A collaborative semantic web layer to enhance legacy systems. *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC'07)*, volume 4825 de *Lecture Notes in Computer Science*, pages 764–777. Springer, 2007.

Workshops et autres conférences avec comité de lecture

- 2011** [↑CO-22] J. Foster, O. Çetinoglu, J. Wagner, J. Le Roux, S. Hogan, J. Nivre, D. Hogan et J. Van Genabith. #hardtoparse : POS Tagging and Parsing the Twitterverse. *Proceedings of the Workshop On Analyzing Microtext (AAAI 2011)*, pages 20–25, États-Unis, 2011. Enterprise Ireland Commercialisation Fund (CFTD/2007/229) and the Science Foundation Ireland (Grant 07/CE/ I1142) as part of the Centre for Next Generation Localisation.
- [↑CO-21] J. Le Roux, B. Favre, S. A. Mirroshandel et A. Nasr. Modèles génératif et discriminant en analyse syntaxique : expériences sur le corpus arboré de Paris 7. *Actes de la 18ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'11)*, pages 371–383, France, 2011.
- [↑CO-20] A. Nasr, F. Béchet, J.-F. Rey, B. Favre et J. Le Roux. MACAON : An NLP Tool Suite for Processing Word Lattices. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT 2011 System Demonstration)*, pages 86–91, États-Unis, 2011.
- 2010** [↑CO-19] M. Attia, J. Foster, D. Hogan, J. Le Roux, L. Tounsi et J. Van Genabith. Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French. *Proceedings of the First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, pages 67–75, États-Unis, 2010. Enterprise Ireland (CFTD/07/229 and PC/09/037) and the Irish Research Council for Science Engineering and Technology (IRCSET).
- [↑CO-18] D. Buscaldi et B. Magnini. Grounding toponyms in an italian local news corpus. *Proceedings of GIR'10 Workshop on Geographical Information Retrieval*, 2010.
- 2009** [↑CO-17] D. Buscaldi, J. M. P. Ortega, P. Rosso, L. A. U. n. López, D. Ferrés et H. Rodríguez. Geotextmess : result fusion with fuzzy borda ranking in geographical information retrieval. *CLEF'08 : Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, pages 867–874, Berlin, Heidelberg, 2009. Springer-Verlag.
- [↑CO-16] D. Buscaldi et P. Rosso. Georeka : Enhancing Web Searches with Geographical Information. *Proc. Italian Symposium on Advanced Database Systems SEBD-2009*, pages 205–212, Camogli, Italy, 2009.
- [↑CO-15] D. Buscaldi et P. Rosso. Indexing with wordnet synonyms may improve retrieval results. *CLEF'09 : Proceedings of the 10th cross-language evaluation forum conference on Multilingual information access evaluation*, pages 128–134, Berlin, Heidelberg, 2009. Springer-Verlag.

- [↑CO-14] D. Buscaldi et P. Rosso. Using GeoWordNet for Geographical Information Retrieval. *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers.*, pages 863–866, 2009.
- [↑CO-13] J. Le Roux. Analyse déductive pour les grammaires d’interaction. *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles (TALN’09)*, pages 1–10, SENLIS, France, 2009.
- [↑CO-12] J. Turmo, P. R. Comas, S. Rosset, O. Galibert, N. Moreau, D. Mostefa, P. Rosso et D. Buscaldi. Overview of qast 2009. *CLEF’09 : Proceedings of the 10th cross-language evaluation forum conference on Multilingual information access evaluation*, pages 197–211, Berlin, Heidelberg, 2009. Springer-Verlag.
- 2008** [↑CO-11] D. Buscaldi et P. Rosso. Map-based vs. knowledge-based toponym disambiguation. *GIR ’08 : Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 19–22, New York, NY, USA, 2008. ACM.
- [↑CO-10] D. Buscaldi et P. Rosso. On the relative importance of toponyms in geoclef. pages 815–822, 2008.
- [↑CO-9] B. Guillaume, J. Le Roux, J. Marchand, G. Perrier, K. Fort et J. Planul. A Toolchain for Grammarians. *Proceedings of the 22nd International Conference on Computational Linguistics, Demo Proceedings (CoLing Demos 2008)*, pages 149–152, Manchester, Royaume-Uni, 2008.
- [↑CO-8] S. Schmitz et J. Le Roux. Calculs d’unification sur les arbres de dérivation TAG. *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles (TALN’08)*, pages 320–329, Avignon, France, 2008.
- [↑CO-7] S. Schmitz et J. Le Roux. Feature Unification in TAG Derivation Trees. C. Gardent et A. Sarkar, éditeurs, *9th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+9)*, pages 141–148, Tübingen, Allemagne, 2008. 12 pages, 4 figures.
- 2007** [↑CO-6] G. Bonfante et J. Le Roux. Intersection Optimization is NP-Complete. *6th International Workshop on Finite-State Methods and Natural Language Processing (FSM/NLP 2007)*, pages 1–14, Postdam, Allemagne, 2007.
- [↑CO-5] D. Buscaldi, Y. Benajiba, P. Rosso et E. Sanchis. Web-based anaphora resolution for the quasar question answering system. *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 324–327, 2007.
- [↑CO-4] D. Buscaldi et P. Rosso. A comparison of methods for the automatic identification of locations in wikipedia. *GIR ’07 : Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 89–92, New York, NY, USA, 2007. ACM.
- [↑CO-3] D. Buscaldi et P. Rosso. Some experiments in humour recognition using the italian wiquote collection. *WILF ’07 : Proceedings of the 7th international workshop on Fuzzy Logic and Applications*, pages 464–468, Berlin, Heidelberg, 2007. Springer-Verlag.
- [↑CO-2] D. Buscaldi et P. Rosso. Upv-wsd : combining different wsd methods by means of fuzzy borda voting. *SemEval ’07 : Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 434–437, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
- [↑CO-1] H. Zargayouna, B. Safar et C. Reynaud. TaxoMap in the OAEI 2007 alignment contest. *Proceedings of the 2d International Workshop on Ontology Matching (OM’07)*, pages 268–275, Busan, Corée, République De, 2007.

Édition d’ouvrages collectifs

- 2010** [↑ED-3] C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari et L. Prévot, éditeurs. *Ontology and the Lexicon, A Natural Language Processing Perspective*. Studies in Natural Language Processing. Cambridge University Press, April 2010, ISBN 978-0-521-88659-8.
- 2008** [↑ED-2] A. Gangemi et J. Euzenat, éditeurs. *Knowledge Engineering : Practice and Patterns, Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW’08)*, volume 5268 de *Lecture Notes in Computer Science*. Springer, 2008, ISBN 978-3-540-87695-3.
- [↑ED-1] A. Gangemi, J. Keizer, V. Presutti et H. Stoermer, éditeurs. *Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives (SWAP2008), Rome, Italy, December 15-17, 2008*, volume 426 de *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

Thèses et Habilitations (TH)

- 2010** [↑TH-2] D. Buscaldi. Toponym disambiguation in information retrieval. Thèse, Universidad Politécnica de Valencia, Valencia, Spain, October 2010. <http://riunet.upv.es/handle/10251/8912>.
- 2007** [↑TH-1] J. Le Roux. La coordination dans les grammaires d'interaction, octobre 2007.

Divers (Div)

[↑Ra-1] G. Bonfante et J. Le Roux. Intersection Optimization is NP Complete, mai 2007. Rapport technique.



Partie III
Acronymes

A³ Apprentissage Artificiel et Applications

AERES Agence d'évaluation de la recherche et de l'enseignement supérieur

AM allocataire moniteur

ANR Agence National de la Recherche

AOC Algorithmes et Optimisation Combinatoire

ATER attaché temporaire d'enseignement et de recherche

CIFRE Conventions Industrielles de Formation par la Recherche

CNR Consiglio Nazionale delle Ricerche

CNRS Centre National de la Recherche Scientifique

CR chargé de recherche

DR directeur de recherche

EA Equipe d'Accueil

EFL Labex *Empirical Foundations of Linguistics*

FEDER Fonds Européen de Développement Régional

FUI Fonds Unique Interministériel

LabEx Laboratoire d'Excellence (appel Grand Emprunt)

IG Institut Galilée

INS2I Institut des Sciences de l'Information et de leurs Interactions, CNRS

INSHS Institut des Sciences Humaines et Sociales, CNRS

Inria Institut National de la Recherche en Informatique et Automatique

IR Ingénieur de Recherche

IRIT Institut de Recherche en Informatique de Toulouse - UMR 5505

IUT Institut Universitaire Technologique

IUTB IUT de Bobigny

IUTV IUT de Villetaneuse

L2TI Laboratoire de traitement et transport de l'information, université Paris 13 - EA 3043

LAGA Laboratoire Analyse, Géométrie et Applications, université Paris 13 - UMR 7539

LATTICE Langues, Textes, Traitements Informatiques, Cognition - UMR 8094

LCR Logique, Calcul et Raisonnement

LIAFA Laboratoire d'Informatique Algorithmique : Fondements et Applications

LIFM Laboratoire d'Informatique Fondamentale de Marseille - UMR 6166

LIPADE Laboratoire d'Informatique Paris Descartes

LIPN Laboratoire d'Informatique de Paris-Nord - UMR 7030

LRI Laboratoire de Recherches en Informatique, université Paris 11 - UMR 8623

MCF maître de conférences

MENESR Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche

AOC Algorithmes et Optimisation Combinatoire

CALIN Combinatoire, algorithmique et interactions

PPF Plan Pluri-Formation

PPS Preuves Programmes et Systèmes - UMR 7126

PRES Pôle de Recherche et d'Enseignement Supérieur

PU Professeur des Universités
RCLN Représentation des Connaissances et Langage Naturel
RNTL Réseau National de recherche et d'innovation en Technologies Logicielles
STIC Sciences et Technologies de l'Information et de la Communication
UMR Unité Mixte de Recherche