Partie I
Activités de recherche

Chapitre 1
RCLN (Représentation des Connaissances et Langage Naturel)

1.1 Membres de l'équipe

(*membres au 11/09/2007*, en italique membres quittant ou ayant quitté le LIPN)

Permanents

Nom	*Prénom*	*Situation*	*Institution*
 Audibert	Laurent	MCF	IUTV
 Després	Sylvie	MCF	IG
 Gayral	Françoise	MCF	IUTV
 Hamon	Thierry	MCF	IG
 Kayser	Daniel	PU	IG

1.1 Membres de l'équipe

	Lévy	François	PU	IUTV
	Nazarenko	Adeline	PU	IG
	Poibeau	Thierry	CR	CNRS
	Recanati	Catherine	MCF	IG
	Rozenknop	Antoine	MCF	IG
	Salotti	Sylvie	MCF	IUTB
	Szulman	Sylvie	MCF	IUTV

Doctorants

Nom	*Prénom*	*Situation*	*Institution*
 Bossard	Aurélien	AM	Université Paris 13
 Bouffier	Amanda	ATER	Université Paris 13
 Boutouhami	Sara	Projet MICRAC	Université Paris 13
 Cailliau	Frederik	CIFRE	Sinequa
 Hamon	Olivier	CIFRE	ELDA
 Messiant	Cédric	AM-DGA	Université Paris 13
 Mondary	Thibault	AM	Université Paris 13
 Weissenbacher	Davy	ATER	Université Paris 13

Impétrants

Nom	*Prénom*	*Situation actuelle*	*Institution*
 *Bourgois*	*Laure*	*Enseignante*	*Ecole d'ingénieur (Dakkar)*
 *Nouioua*	*Farid*	*Post-doctorant*	*LRI*

Personnel temporaire du 1/9/2003 au 1/10/2007

Nom	*Prénom*	*Situation*	*Début*	*Fin*	*Financement*
 *Aubin*	*Sophie*	*Ingénieur de recherche*	*2/2004*	*12/2006*	*Contrat de recherche*
 Ceausu-Dragos	Valentina	ATER	2/2007		Contrat de recherche
 *Derivière*	*Julien*	*Ing. de recherche*	*5/2004*	*12/2006*	*Contrat de recherche*
 *Jacques*	*Marie-Paule*	*Post-doctorante*	*10/2006*	*08/2007*	*Contrat de recherche*
 *Vauvert*	*Guillaume*	*Post-doctorant*	*3/2004*	*8/2004*	*Contrat de recherche*
 *Zargayouna*	*Haifa*	*Post-doctorante*	*5/2006*	*11/2006*	*Contrat de recherche*
 *Zimina*	*Maria*	*Post-doctorante*	*5/2006*	*11/2006*	*Contrat de recherche*

Personnels invités du 1/9/2002 au 1/9/2006 (au moins 1 semaine)

Prénom	*Nom*	*Origine*	*Période*	*Financement*
Sofia	Ananiadou	Manchester Univ. & NaCTem (UK)	5-28 juin 2006	professseur invité
Anna	Korhonen	Cambridge Univ., Cambridge (UK)	15-30 mai puis 1-15 juillet 2007	rofessseur invité
Aïcha	Mokhtari	USTHB, Alger (DZ)	5-20 juillet 2004	Projet CMEP-CNRS
Aïcha	Mokhtari	USTHB, Alger (DZ)	10-30 juillet 2005	Projet CMEP-CNRS

1.1.1 Evolution de l'équipe

• Permanents

Départs

- B. Biébow, MCF (thème ingénierie des connaissances), départ à la retraite (2005)

Promotion

- A. Nazarenko, promue de MCF à PU (thème traitement automatique des langues) (2005)

Arrivées

- L. Audibert, MCF IUTV (thème traitement automatique des langues), 2004 (thèse Aix-Marseille)
- A. Rozenknop, MCF IG (thème traitement automatique des langues), 2005 (thèse EPFL, Suisse)
- S. Després, MCF IG (ingénierie des connaissances textuelles), 2007 (mutation)

• Doctorants

Trois thèses ont été soutenues sur la période :

- Touria Aït El Mekki a terminé sa thèse de doctorat sur la « Construction semi-automatique d'index de fin de livres », soutenue le 21/12/2004 avec mention Très Honorable. Elle est maître de conférences à l'Université d'Angers depuis sept. 2005.
- Farid Nouioua (Dir. D. Kayser) a soutenu en avril 2007 sa thèse de doctorat intitulée « Extraction et Utilisation des Normes pour un Raisonnement Causal dans un Corpus Textuel »). Il est actuellement post-doctorant au LRI (Université Paris-Sud).
- Laure Bourgois (Dir. F. Lévy) a soutenu en juin 2007 sa thèse de doctorat sur la « Représentation et comparaison de Web services complexes avec des logiques dynamiques ». Elle est actuellement en poste dans une école d'ingénieurs à Dakkar.

Six étudiants sont en cours de thèse au sein de l'équipe RCLN :

- Davy Weissenbacher s'intéresse au problème de l'incomplétude et de l'imperfection des informations entrant dans la résolution de l'anaphore pronominale depuis 2003 (Dir. C. Fouqueré et A. Nazarenko). La soutenance est prévue fin 2007.

- Amanda Bouffier a commencé fin 2004 une thèse sur l'analyse de structures discursives au sein de documents procéduraux, afin d'aider leur modélisation (Dir. D. Kayser et T. Poibeau). Elle est actuellement ATER.
- Aurélien Bossard a commencé fin 2006 une thèse comme allocataire-moniteur. Il s'intéresse à l'évolution et la fusion d'informations extraites de données textuelles (Dir. D. Kayser et T. Poibeau).
- Cédric Messiant a débuté une thèse fin 2006 sur l'acquisition automatique d'informations linguistiques sur corpus. Il est financé par une bourse DGA (Dir. A. Nazarenko et T. Poibeau).
- Thibault Mondary, qui est en thèse depuis fin 2006 comme allocataire-moniteur, s'intéresse à la construction d'ontologies à partir de textes (Dir. A. Nazarenko et S. Després).
- Sara Boutouhami a commencé fin 2006 une thèse sur l'argumentation dans la langue naturelle (Dir. D. Kayser) qui est financée dans le cadre d'un projet ANR.

Deux doctorants de l'équipe font actuellement leur thèse dans l'industrie :

- Frederik Cailliau est en thèse chez Sinequa depuis 2004 (Bourse Cifre - Dir. A. Nazarenko). Il travaille à la conception et à la mise en oeuvre d'un système de gestion de ressources linguistiques.
- Olivier Hamon est en thèse à ELDA depuis 2005 (Bourse Cifre - Dir. A. Nazarenko). Il définit des méthologies et des protocoles d'évaluation des technologies de la langue et de la communication homme-machine.

Bilan

Sur la période, l'équipe a augmenté son potentiel de recherche en traitement automatique des langues sur le thème « sémantique de corpus » (arrivée de Laurent Audibert en 2004 après celle de Thierry Poibeau en 2003), ce qui a permis de créer un pôle fort sur ce thème et de participer à des projets ambitieux, tout en renforçant le lien avec le thème « compréhension fine de textes ». L'arrivée d'Antoine Rozenknop en 2005 est venu compléter ce pôle en lui apportant des compétences en analyse probabiliste, qui faisaient défaut.

Sylvie Després (mutation 2007) est venue renforcer le thème « ingénierie des connaissances textuelles » qui reste cependant faible numériquement.

Après quelques années creuses en 2003-2004, le nombre de thèses encadrées par l'équipe a sensiblement augmenté.

A signaler également le recrutement sur la période 2004-2007 de six ingénieurs de recherche ou postdoctorants financés sur des contrats de recherche (S. Aubin, J. Derivière, G. Vauvert, M. Zimina, M.-P. Jacques, H. Zargayouna), ce qui a permis à l'équipe de poursuivre ses travaux de recherche théorique tout en se lançant dans des tâches de développement. Ces derniers permettent de valoriser les résultats des recherches antérieures et de se doter d'outils pour mener de nouvelles expérimentations.

1.2 Orientation scientifique générale

Les connaissances sont très généralement transmises par l'intermédiaire du langage, ce qui justifie que la Représentation des Connaissances, enjeu majeur de l'Intelligence Artificielle et des Sciences Cognitives, ait pour objectif de traiter les connaissances exprimées en Langage Naturel. De manière complémentaire, la compréhension de textes met en oeuvre des connaissances et des raisonnements qu'il est fondamental de décrire et modéliser.

L'équipe s'intéresse donc au langage, non pour ses propriétés en tant que système formel de signes, mais pour son pouvoir expressif, et à la Représentation des Connaissances, notamment en tant qu'outil mis au service du traitement du Langage Naturel.

Les travaux de l'équipe comportent à la fois des recherches à caractère fondamental et des recherches plus appliquées. L'évolution des thèmes de recherche de l'équipe a mené à travailler dans trois directions principales :

Compréhension fine de textes Il s'agit étudier finement, pour le modéliser, le calcul sémantique entrant dans certains phénomènes interprétatifs (liés au pluriel, au temps, aux normes).

Sémantique de corpus Ce thème vise à développer des outils destinés à faciliter l'accès au contenu de documents techniques ou de vastes bases textuelles, comme le Web,

Ingénierie des connaissances textuelles Ce troisième thème a pour objectif de développer des outils d'ingénierie des connaissances à partir de textes.

Ces activités sont détaillées ci-dessous comme trois thèmes de recherche distincts. En réalité, elles sont fortement liées. L'objectif commun est de développer des méthodes permettant de modéliser, à des degrés divers, le contenu des textes et documents, soit pour y faciliter la navigation et la recherche d'information, soit pour en restituer l'information factuelle, soit encore pour construire un modèle du domaine. Les recherches plus fondamentales du thème « compréhension fine de textes » nourrissent la réflexion du thème « sémantique de corpus », même si la nécessaire robustesse des traitements sur corpus imposent des simplifications. Le thème « ingénierie des connaissances à partir de textes » exploite les résultats des deux premiers thèmes pour élaborer des ontologies fondées sur les connaissances textuelles. Ces thèmes sont également liés à ceux de l'équipe A3 pour tout ce qui concerne l'apprentissage à partir de données textuelles.

Ces questions sont abordées à travers l'étude de différents problèmes applicatifs et dans le cadre de projets collaboratifs variés. La diversité des domaines et des corpus étudiés permet de tester nos hypothèses de travail et de mettre en oeuvre nos méthodes sur des problèmes concrets tout en testant leur robustesse. Nous avons étudié plus particulièrement les problématiques suivantes :

Corpus de constats d'accident L'équipe RCLN poursuit ses recherches sur la compréhension des constats d'accident, notamment pour ce qui concerne l'analyse des normes qui sont implicites dans les textes mais qui en conditionnent l'interprétation (projet MICRAC).

Corpus de génomique Nous poursuivons également notre travail d'analyse de la bibliographie en génomique : il s'agit à la fois d'extraire des informations sur les interactions géniques (projet ExtraPloDocs) et et de développer un moteur de recherche spécialisé (projet ALVIS).

Documents médicaux L'ensemble de l'équipe s'est engagé dans la conception de méthodes et d'outils d'accès au contenu des documents médicaux, depuis deux ans et à la faveur du projet ACTMed. Ce projet est un Programme Pluriformations commun à trois laboratoires de l'Université Paris 13[1]. Il a été lancé en 2005 pour quatre ans. L'ambition de ce projet est de créer un pôle d'expertise rassemblant des compétences en traitement automatique des langues, ingénierie des connaissances, analyse de l'information médicale et linguistique. Ce projet permet de fédérer et de concrétiser des résultats obtenus à travers les trois thèmes de recherche de l'équipe RCLN.

Corpus de directives européennes Depuis deux ans, nous avons commencé à travailler sur le domaine juridique pour construction des ressources juridique, à la faveur de notre participation au projet LOIS[2]. Nous souhaitons développer ce nouveau champ de recherche dans les prochaines années, notamment au travers de collaborations avec des partenaires européens.

1.3 Thèmes de recherche

1.3.1 Compréhension fine des textes

Ce thème s'attache à rendre compte d'une compréhension fine des textes qui s'oppose aux analyses nécessairement plus simples requises pour un traitement de corpus et qui se définit pour nous comme la capacité à tirer du texte des inférences adéquates.

Dans cet objectif, et à travers l'étude de certains phénomènes interprétatifs complexes, comme la référence aux événements, le pluriel ou les normes, nous examinons la dynamique du processus interprétatif et les principes sur lesquels il repose. Nous en dégageons plusieurs idées fortes :

[1] Le LIPN, le Laboratoire de Linguistique-Informatique de Paris 13 (LLI) et le laboratoire d'Informatique Médicale et Bioinformatique (Lim&Bio)

[2] Le LIPN à été associé à ce projet, mais sans financement.

- L'hypothèse de compositionnalité, généralement retenue pour le calcul du sens des expressions complexes, se heurte à des difficultés majeures et il est préférable d'envisager d'autres formes de calcul sémantique plus plausibles cognitivement [CI-6].
- Ce processus interprétatif mobilise des connaissances de différents niveaux (sémantiques : lexicales et grammaticales, connaissances liées au domaine, au type du textes) qu'il faut représenter et dont il faut examiner les interactions.
- La prise en compte de l'incertain, la non monotonie du raisonnement et son caractère analogique sont essentiels.

• Comprendre les relations temporelles

Nous nous intéressons à la compréhension des relations temporelles entre événements, qu'elles soient marquées syntaxiquement ou implicites. On peut à l'heure actuelle contraster la finesse des analyses linguistiques (sémantique lexicale et sémantique des événements, notamment) qui distinguent la position de l'événement sur l'axe des temps (passé, présent, futur), sa dynamique interne (état, événement, processus), son aspect (duratif ou ponctuel, borné ou indéfini, continu ou itératif) et une variété de relations de succession (succession immédiate ou distante, transitions progressives ou brutales), avec le caractère plus rudimentaire des représentations calculables (essentiellement l'axe des réels ou des entiers, ou les relations d'une algèbre d'intervalles).

C'est à cette complexité de la représentation des événements que s'est intéressé le projet dirigé par François Lévy dans le programme interdisciplinaire TCAN du CNRS. L'accent à été mis sur la reconnaissance d'événements itératifs et l'analyse/représentation de leur aspect. S'agissant d'événements simples (non itératifs), l'extension proposée par L. Gosselin de la théorie de Reichenbach à des intervalles permet de rendre compte assez finement de l'aspect avec des règles automatisables : à chaque événement est attaché un intervalle d'énonciation, un intervalle de procès et un intervalle de référence (celui sur lequel l'attention est attirée, et qui peut ne recouvrir que partiellement ou même pas du tout le procès). L'étude menée dans le groupe a montré que l'itération met l'accent diversement sur les séries d'intervalles que l'on peut construire, en focalisant soit sur la série en tant que telle, soit sur l'intervalle global la contenant. On peut en rendre compte en combinant des constructeurs d'intervalles complexes. Les difficultés rencontrées sont liées à la délimitation des cadres de discours. Un corpus de 18 000 exemples de phrases itératives extraites du journal *Le Monde* a été réalisé.

Par ailleurs, une étude sur le temps et l'aspect dans le corpus des constats d'accident se poursuit. Catherine Recanati, en collaboration avec l'équipe A^3, explore la possibilité de faire apparaître des séquences de motifs verbaux, participant d'un niveau discursif de granularité frontière entre syntaxe et sémantique (les étiquettes sont par exemple le temps grammatical doublé de la catégorie aspectuelle lexicale du verbe). Les algorithmes utilisés relèvent de l'apprentissage numérique non supervisé (connexioniste et stochastique).

• Réexaminer le concept de cause, extraire des normes

Nous avons entrepris depuis quelques années de réexaminer le concept de cause. En effet, il nous semble que mettre à la base d'une théorie sémantique les notions de référence et de conditions de vérité n'est pas satisfaisant pour rendre compte du processus de compréhension. En revanche, beaucoup plus que la dichotomie du vrai et du faux, la compréhension semble utiliser le contraste entre le normal et l'anormal, et ceci est particulièrement visible lorsqu'on analyse l'usage courant de la causalité.

Daniel Kayser et Farid Nouioua ont repris le travail sur un corpus de constats d'accidents qui avait fait l'objet d'études antérieures au sein de l'équipe. Il apparaît que, parmi les causes d'accident qui peuvent être données (p.ex. la présence d'essence dans le réservoir est une condition nécessaire pour qu'une voiture roule, donc pour qu'elle dérape), seules celles qui sont perçues comme des anomalies (la présence d'une flaque d'huile sur la chaussée) sont effectivement pertinentes [CI-4]. D. Kayser et F. Nouioua essayent donc de rendre compte du raisonnement tenu par le lecteur pour trouver la cause de l'accident, au moyen de règles d'inférence non-monotones (voir ci-dessous) [RI-4]. La méthode sera validée si, sur un échantillon différent de celui ayant servi à faire l'analyse, le système identifie la cause de l'accident comme l'aurait fait un lecteur humain.

Cette méthode d'extraction des normes peut avoir des conséquences importantes en sémantico-pragmatique, car les inférences tirées de normes sont beaucoup plus riches que les inférences sémantiques habituelles, fondées sur la préservation de la vérité.

Une première synthèse de cette approche de la causalité, ainsi qu'une étude concernant les réseaux bayésiens, a été publiée [RE-5] dans un numéro spécial de revue sur le raisonnement causal, numéro dont le LIPN a pris l'initiative et qu'il a piloté.

Ce travail se prolonge actuellement dans le cadre d'un projet inter-disciplinaire soutenu par l'Agence Nationale de la Recherche (projet MICRAC : Modèles Informatiques et Cognitifs du RAisonnement Causal) qui associe un Laboratoire de Psychologie (Toulouse Le Mirail) et trois Laboratoires d'Informatique (IRIT, CRIL, LIPN) pour recenser différentes approches philosophiques, logiques, et informatiques de la causalité et tester celles qui correspondent le mieux au raisonnement causal humain.

• Interpréter le pluriel

L'étude de l'interprétation sémantique du pluriel que nous avons réalisée à partir d'exemples extraits d'articles de journaux nous a conduits au constat suivant : les problèmes sémantiques soulevés sont beaucoup plus divers et riches que ceux évoqués par la littérature traditionnelle traitant de ce sujet, loin de la seule dichotomie entre lecture distributive et lecture collective généralement relevée.

Nous avons mis en évidence le rôle du temps qui intervient de façon essentielle à la fois dans la construction des collections associées aux expressions plurielles et dans leur évolution. En effet, la compréhension de nombreux énoncés oblige les expressions plurielles à être considérées à différents instants. Françoise Gayral, Daniel Kayser et François Lévy ont montré que deux interprétations sont possibles [CO-17] : une interprétation que nous appelons *de re* où la collection est composée d'éléments fixes même si chaque élément a des propriétés qui évoluent dans le temps (énoncé 1 ci-dessous) et une interprétation *de dicto* où la collection persiste alors que ses éléments changent. Dans l'énoncé 2 en effet, l'expression "les bacheliers" convoque deux réalisations distinctes de la même collection *de dicto* (les bacheliers de l'an dernier et les bacheliers de cette année). F. Gayral, D.Kayser et F. Lévy ont également étudié les différents cas où une partition de la collection en sous-ensembles était indispensable à l'interprétation de la phrase [CO-3].

1. *Les bacheliers de ma classe sont maintenant inscrits à l'université.*
2. *Les bacheliers sont moins nombreux cette année que l'an dernier.*
3. *Moins nombreux, les étudiants proviennent de plus en plus d'Europe et de moins en moins d'Afrique.*

F. Gayral, D.Kayser et F. Lévy ont par ailleurs montré comment les trois questions ci dessus (caractère *de re/de dicto* des collections, existence de partitions et critère sur lequel s'effectue le partitionnement, caractère distributif ou collectif de la (des) prédication(s) s'appliquant à l'expression plurielle) sont extrêmement liées et mobilisent des connaissances de différents niveaux. L'étude de nombreux énoncés attestés a montré qu'il est impossible de ne s'intéresser qu'à une de ces questions, de l'isoler et d'y répondre sans faire intervenir la réponse aux autres. Dans l'énoncé 3 par exemple, l'interprétation du seul pluriel "les étudiants" donne lieu à la création d'au moins six ensembles et à différentes prédications concernant ces étudiants, prédication collective pour leur dénombrement, distributive pour leur provenance.

• Représenter les procédures interprétatives

Dans beaucoup d'approches en linguistique computationnelle ou en sémantique formelle, interpréter une phrase est vu comme l'application d'un mécanisme récursif combinant des informations lexicales et syntaxiques. L'hypothèse sous-jacente est l'hypothèse de compositionnalité qui postule que le sens d'une expression complexe est déterminé par sa structure et par le sens de ses constituants.

Françoise Gayral, Daniel Kayser et François Lévy développent trois critiques majeures pour montrer que cette hypothèse est inapte à rendre compte de l'activité de compréhension [CL-6].

- En considérant la polysémie comme une énumération de sens, elle oblige à considérer le processus interprétatif comme un processus de désambiguisation parmi des listes prédéfinies de sens alors que de nombreux glissements de sens difficilement prévisibles peuvent intervenir en contexte.
- par le « localisme » et l'incrémentalité du processus qu'elle implique, elle ne fait dépendre le sens d'une expression complexe que du sens de ses constituants et de sa structure au détriment d'autres facteurs dont nous montrons au contraire l'importance.
- par le but qu'elle fixe au processus interprétatif (construire une représentation unique du sens de l'expression et donc de chacun de ses composants), elle empêche de rendre compte du phénomène de « coprésence » dans lequel un même mot peut prendre des sens différents suivant la partie de la phrase auquel on le rattache alors que cette coprésence est courante dans les textes.

A partir de cette critique, nous proposons une autre approche basée sur une logique non-monotone dans laquelle nous tentons d'exprimer les règles d'interprétation [CI-8]. Nous avons choisi la logique des défauts de Reiter par les avantages qu'elle procure. Par la capacité à exprimer des règles de défaut, ce formalisme permet de rétablir la notion d'exception dans la position importante qu'elle a en grammaire, position d'où les théories purement compositionnelles l'avaient évincée. Il autorise la non-monotonie du raisonnement, c'est-à-dire la remise en cause de conclusions jusque-là acquises. Grâce à la notion d'extension définie comme la solution d'une équation de point fixe, il rend compte de la circularité liée à l'influence du contexte sur l'interprétation d'un mot, sachant que le mot influe lui aussi sur l'interprétation de son contexte. En contrepartie, le principal problème de ce type d'approches est celui du contrôle : il est absolument nécessaire de bloquer la dérivation de certaines valeurs sémantiques si celles-ci s'avèrent inadéquates et, à l'inverse, d'en privilégier d'autres [RI-1].

En marge de ces travaux sur la logique des défauts, nous explorons d'autres voies pour modéliser la prise en compte de l'incertain dans les processus interprétatifs. Dans le cadre de sa thèse sur la résolution d'anaphore qui est un problème de traitement automatique épineux, Davy Weissenbacher montre que les modèles graphiques probabilistes comme les réseaux bayésiens [RE-13*] sont adaptés pour modéliser un faisceau de connaissances à la fois incertaines, partielles et interdépendantes et les exploiter pour identifier l'antécédent d'un pronom anaphorique.

Nous nous intéressons par ailleurs aux capacités « analogiques » mises en oeuvre dans la compréhension du discours et dans le raisonnement humain. Plusieurs psychologues et philosophes ont souligné en effet de manière convaincante le caractère analogique de nos représentations cognitives. Pour élucider cette notion, C. Recanati s'est attachée à caractériser formellement la distinction texte/diagramme dans les raisonnements [RE-10].

1.3.2 Sémantique de corpus

En parallèle et en complément avec les analyses fines développées dans le cadre du thème précédent, nous explorons des méthodes d'analyses sémantiques adaptées au traitement de grandes quantités de texte.

Au-delà des techniques de recherche d'information qui associent un ensemble de documents à une requête, un enjeu majeur aujourd'hui concerne l'accès au contenu même des documents textuels. Il s'agit à la fois d'exploiter des documentations scientifiques et techniques et de maîtriser le contenu sémantique des pages web par l'ajout d'+s sémantiques interprétables par une machine (web sémantique) [CL-3]. Dans tous les cas, accéder au contenu des documents repose sur une analyse sémantique partielle et robuste du corpus.

Ces travaux s'inscrivent dans un courant de recherche important qui se focalise sur les moteurs de recherche sémantiques, les techniques d'extraction d'information et de question-réponse. Notre approche de ces problèmes est originale à plusieurs égards : nous mettons l'accent sur l'analyse de corpus spécialisés plutôt que sur des textes tout-venant ; nous articulons une analyse linguistique riche avec des techniques d'apprentissage et des méthodes d'acquisition de ressources ; nous cherchons à dépasser les frontières entre les techniques traditionnelles (recherche d'information, extraction d'information, question-réponse, navigation hypertextuelle) pour élaborer des approches plus riches de compréhension de textes.

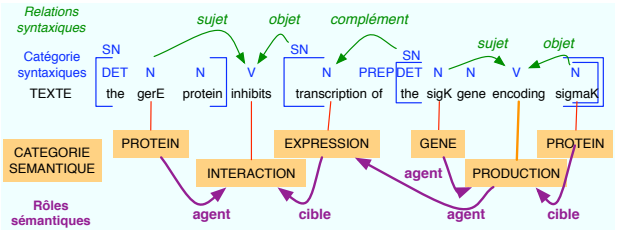


FIG. 1.1: Exemple de texte enrichi d'annotations linguistiques. Les différents niveaux d'annotation correspondent à des description de plus en plus abstraites et normalisées du matériau textuel.

- **Concevoir de nouvelles méthodes d'accès au contenu textuel**

En nous appuyant sur des outils d'analyse de corpus (voir plus loin), nous développons des méthodes facilitant l'accès au contenu textuel par le lecteur. Nous explorons en parallèle différentes approches, parfois sur les mêmes données, l'idée étant à terme d'arriver à combiner dynamiquement les différentes méthodes d'accès au contenu textuel [CL-5] :

Extraction d'information en biologie Cette tâche consiste, dans un domaine restreint, à extraire des éléments d'information précis à partir d'un ensemble de textes homogènes et à remplir des formulaires prédéfinis avec ces éléments d'information. Elle vise soit à alimenter des bases de données à partir d'informations extraites des textes, soit à ajouter des annotations sémantiques explicites aux documents. L'équipe RCLN travaille sur ces problème d'extraction depuis 1998 [LI-1], et plus particulièrement dans le domaine de la biologie depuis 2001. Notre participation à différents projets d'extraction d'information (ex. ExtraPloDocs) a montré le coût prohibitif de la tâche d'acquisition des ressources et notamment des règles d'extraction qui guident le processus d'extraction [CL-4]. Thierry Hamon, Adeline Nazarenko, Thierry Poibeau ont mis au point une méthode pour apprendre automatiquement ces ressources à partir de corpus (voir figure **??**) en collaboration avec Erick Alphonse (équipe A3) et des spécialistes de l'unité MIG de l'INRA [CO-19]. Les résultats obtenus montrent l'intérêt des méthodes d'apprentissage relationnel et l'importance de la normalisation linguistique préalable pour l'apprentissage des règles d'extraction. Cette normalisation permet de représenter les données linguistiques en faisant abstraction des variations paraphrastiques de surface (voir figure ci-dessous), ce qui met en évidence des régularités sémantiques et permet l'apprentissage de règles d'extraction.

Recherche d'information spécialisée Nos travaux dans cette direction sont plus récents. Dans le cadre du projet ALVIS, le moteur de recherche spécialisé en biologie exploite les résultats de la plate-forme d'annotation présentée ci-dessous, adaptée à la biologie [CO-38]. Nous cherchons tout d'abord à mesurer l'intérêt des annotations linguistiques et sémantiques pour l'indexation des documents et pour la visualisation des informations dans l'interface utilisateur (si 'cat' est un nom de gène, il est important qu'il soit reconnu comme tel par un étiqueteur d'entités nommées pour éviter qu'il ne soit confondu avec l'animal 'chat' par un moteur de recherche dédié à la biologie). Les expériences correspondantes sont en cours. Sylvie Salotti et Haïfa Zargayouna étudient par ailleurs ce que la prise en compte de la structure même du document (structure XML) apporte à la recherche d'information. Dans le système qui a été développé par Haïfa Zargayouna, les termes sont désambiguïsés en fonction de leur contexte d'apparition [CO-9], lequel est défini par les éléments de structure des documents. Ce système permet d'effectuer une recherche ciblée. Au-delà du document, il donne accès directement aux points d'entrée pertinents dans le document.

Interrogation de documents Poser des questions et avoir directement la réponse sans avoir à lire des documents est une ambition ancienne du traitement automatique des langues. C'est ce que visent les systèmes de question-réponse. Nous avons mis l'accent sur les systèmes spécialisés. En collaboration avec Pierre Zweigenbaum (équipe INSERM U729, désormais

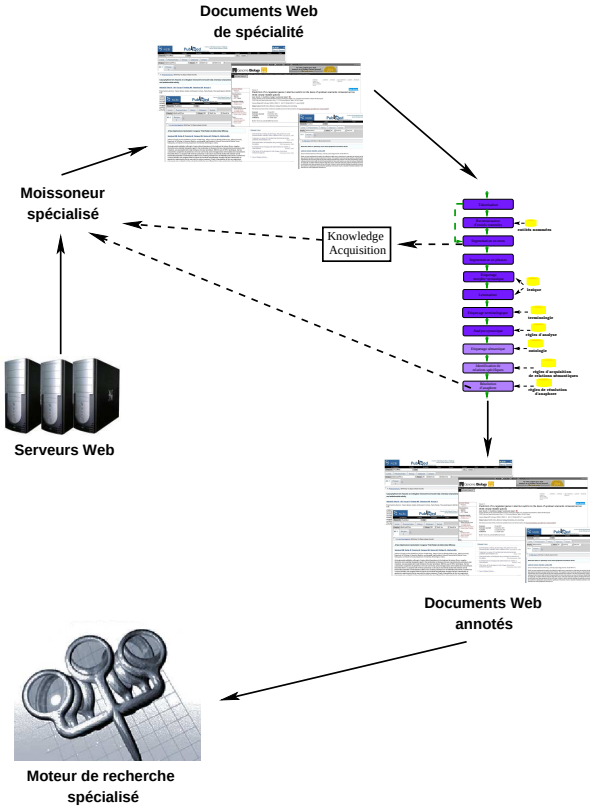


FIG. 1.2: Architecture de la plate-forme d'annotation de documents

LIMSI), Thierry Poibeau a développé un système de question-réponse médical et a participé à la campagne d'évaluation Technolangue EqueR (2004). Celle-ci a montré l'intérêt de la problématique des questions-réponses dans un domaine de spécialité. Nous pensons par ailleurs qu'exploiter la structure des documents permet de mieux guider l'interrogation. Sylvie Salotti et Haïfa Zargayouna ont défini un langage de requête par l'exemple où l'utilisateur peut spécifier les éléments de structure ainsi que le contenu dans un format XML et Victor Rosas montre comment traduire des requêtes en langue naturelle dans ce langage structuré. Ces travaux s'inscrivent dans cadre de la campagne 2006 d'Evaluation INEX (Initiative for the evaluation of XML Retrieval) mais ces travaux sont également testés en interne dans le cadre du projet ACTmed pour l'interrogation par les médecins des guides de bonne pratique médicale. La coopérativité des systèmes de question-réponse est aussi important pour leur meilleure utilisation et conviviabilité. Si la réponse exacte à une question n'est pas présente dans les textes, il peut être intéressant de proposer des réponses relâchant certaines contraintes de la question. Cette question de la coopérativité dans les systèmes de question-réponse a été abordée par Françoise Gayral en collaboration avec Patrick Saint-Dizier.

Navigation documentaire Les méthodes d'analyse de corpus développées au sein de l'équipe RCLN, notamment d'analyse terminologique, nous ont permis de mettre au point des techniques originales d'accès au contenu dans les documents qui s'inspirent des approches traditionnelles en matière de consultation et de navigation documentaire. Le travail de thèse de Touria Aït El Mekki a conduit au développement d'IndDoc, un outil qui aide l'indexeur à produire un index de fin de livre [RE-9] [RI-3]. Ce travail se poursuit aujourd'hui en collaboration avec Laurent Audibert et Haïfa Zargayouna. L'originalité de IndDoc par rapport à d'autres outils d'indexation tient à la qualité de l'analyse du contenu documentaire qui apporte une aide substantielle à l'indexeur pour la sélection de l'information. Cette approche a fait ses preuves dans la construction de l'index de la version électronique d'un ouvrage scientifique (voir figure ci-dessous)et elle suscite aujourd'hui l'intérêt des indexeurs professionnels [RE-12]. Dans le prolongement de ces premiers travaux, Amanda Bouffier et Adeline Nazarenko ont proposé une méthode de navigation qui s'appuie sur une analyse terminologique des documents eux-mêmes, sur leur table des matières et sur un modèle du domaine. L'outil de navigation interactif qui a été réalisé par Thibault Mondary est testé dans le cadre du projet ACTMed pour faciliter la consultation des Guides de Bonne Pratique Clinique (GBP) par les médecins.

• Développer des outils d'analyse de corpus

Ces méthodes d'accès au contenu textuel s'appuient sur des outils d'analyse de corpus performants. Un effort important a été entrepris au sein de l'équipe RCLN pour développer des outils d'analyse de corpus qui permettent d'explorer de nouvelles approches sémantiques, d'analyser le contenu textuel ou d'amorcer l'acquisition de connaissances. Différents outils ont été réalisés à la faveur des projets dans lequels l'équipe RCLN a été impliquée depuis 2002 :

- une plate-forme d'annotation linguistique a été conçue et développée par Thierry Hamon, secondé par Julien Derivière (ingénieur de recherche contractuel) [Lo-6][3]. Cette plateforme d'annotation linguistique peut produire différents types d'annotation, sur le français et l'anglais. Elle intègre les modules ci-dessous.
- divers modules d'analyse ont été réalisés : étiquetage des entités nommées (TagEN, Jean-François Berroyer et Thierry Poibeau), étiquetage terminologique (Thierry Hamon), analyse syntaxique spécialisée pour la biologie (BioLG, Sophie Aubin) [CI-13], désambiguïsation sémantique (Laurent Audibert), résolution d'anaphores (Davy Weissenbacher).

Nous avons mis l'accent sur la robustesse des traitements ainsi que sur la rapidité de l'annotation en proposant une architecture modulaire et distribuée. Thierry Hamon et Julien Derivière ont montré l'efficacité de cette approche lors d'une première expérience menée sur une collection de 55 329 documents en rapport avec la biologie et issus du web (soit plus de 80 millions de

[3]Cette plateforme d'annotation linguistique, développée en Perl, est disponible sous forme de module CPAN (Comprehensive Perl Archive Network – répertoire des modules Perl développés dans le monde) à l'adresse suivante : http ://search.cpan.org/ thhamon/Alvis-NLPPlatform-0.1/.

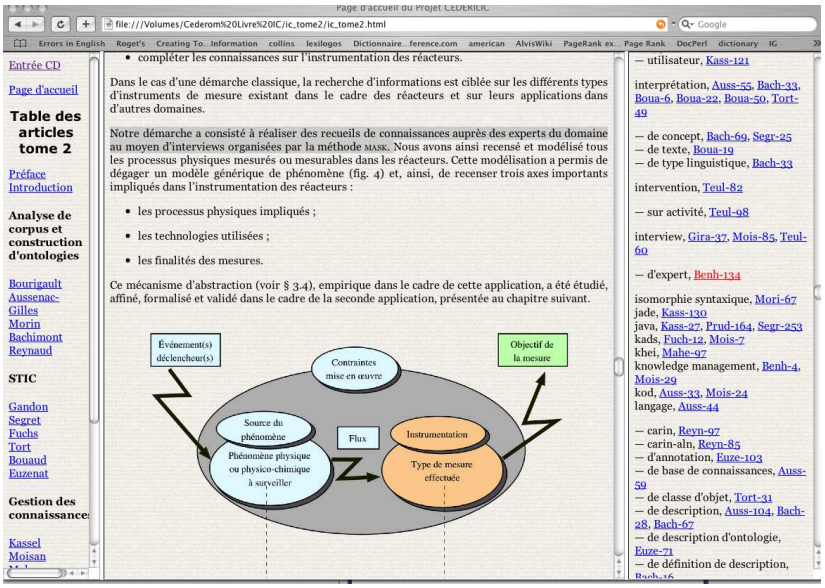


FIG. 1.3: Version électronique de l'index de l'ouvrage *Ingénierie des connaissances*, coordonné par R. Teulier *et al.*, L'Harmattan, 2005. L'écran est divisé en trois zones : la table des matières à gauche, l'index à droite et une fenêtre de navigation dans le document. Le lecteur a consulté l'index et à l'entrée, « interview d'expert », il a cliqué sur le lien « Benh-134» (indiqué en rouge), ce qui provoqué l'affichage du passage correspondant dans la fenêtre de navigation. Cette interface repose sur une indexation sous-jacente du document qui a été réalisée par IndDoc.

mots) [CI-19]. Cette expérience a prouvé que, même s'il reste coûteux, ce processus d'annotation permet de gérer le flux de documents qu'un moteur de recherche spécialisé doit pouvoir indexer. Elle a également montré que notre plate-forme est plus robuste (par rapport au nombre de documents à traiter, à leur taille, à la qualité du nettoyage préalable des données) que les plates-forme concurrentes comme celle de GATE.

Cette plate-forme est conçue pour être adaptée, sans chute de performance, à différents domaines ou à de nouveaux corpus par un processus de spécialisation. Cette spécialisation s'opère de deux manières :

- Par l'exploitation de connaissances spécialisées (dictionnaires d'entités nommées, terminologie, ontologie, notamment) : nous avons montré que l'ambiguïté de l'analyse syntaxique est réduite de 40% [4] quand les connaissances terminologiques peuvent être prises en compte [CI-10] ;
- Par le recours à des méthodes d'apprentissage à partir de corpus. Pour la désambiguïsation sémantique, Laurent Audibert a montré que l'impact de la qualité des indices fournis aux algorithmes d'apprentissage est bien plus important que la qualité intrinsèque des algorithmes employés. Les paramètres du module de résolution des anaphores peuvent eux aussi être adaptés pour un nouveau corpus par apprentissage : Davy Weissenbacher a modélisé le problème de la résolution des anaphores sous la forme d'un réseau bayésien dont les différents paramètres sont appris à partir des fréquences observées dans un corpus étiqueté [CI-12].

En marge de ces travaux qui reposent sur l'analyse linguistique et pour répondre aux besoins internes de nos travaux de recherche, L. Audibert a également développé des outils de manipulation et d'exploration de corpus : Concordanta, XExplora et Explora.

[4] L'ambiguïté est mesurée à partir du nombre d'analyses différentes produites par phrase.

- **Analyser la structure des documents textuels**

Nos travaux sur l'accès au contenu textuel ont fait émerger une nouvelle thématique au sein de l'équipe RCLN. Il nous est apparu que la structure des documents est de première importance pour aider le lecteur à accéder à la bonne information, alors que l'accès à la documentation électronique s'est développé autour de méthodes fondées sur l'analyse de mots clés ou de segments isolés de textes, qui ignorent largement la structure des documents.

Partant de ce constat, nous avons commencé à travailler sur l'analyse de la structure des documents. Il s'agit à la fois d'exploiter les documents semi-structurés (en HTML ou en XML par exemple) pour compléter l'information portée par le contenu textuel et, inversement, d'analyser le texte lui-même pour découvrir des séquences porteuses de sens (titres, blocs informationnels cohérents, etc) et les mettre en relation.

A travers l'étude des Guides de Bonnes Pratiques médicales, choisis au sein du projet ActMed comme corpus de référence, Amanda Bouffier et Thierry Poibeau ont mis en évidence des structures sémantiques qui s'étendent au-delà de la phrase, comme des enchaînement « conditions-actions" et qui sont essentielles pour la lecture et la compréhension de ces documents. L'analyse de ces structures pose des problèmes linguistiques fondamentaux : comment repérer ces séquences de textes ? A partir de quelle combinaison d'indices ? Quelle est la portée d'une condition ? Le travail de thèse d'A. Bouffier vise à répondre à ces questions, à partir d'une analyse fine de documents médicaux. Il sera étendu afin de s'assurer de la généricité de l'approche [CO-43].

Ces recherches sont complétées par l'analyse fine d'autres éléments pertinents comme les titres. En fonction de leur contenu, nous cherchons à les typer selon une ontologie du domaine. Il s'agit d'aider à l'interprétation du document afin d'offrir des parcours de lecture pertinents et non linéaires, lors du passage au document électronique.

La prise en compte de la structure des documents doit permettre d'offrir de nouveaux moyens d'accès aux textes et d'enrichir les outils d'accès au contenu présentés ci-dessus (outils de navigation, notamment).

1.3.3 Ingénierie des connaissances textuelles

Le troisième volet des recherches de l'équipe RCLN concerne l'acquisition de connaissances à partir de textes. Le développement de l'analyse de corpus a montré en effet qu'on peut s'appuyer sur le matériau textuel pour amorcer et guider le processus complexe d'acquisition qui vise à élaborer les ressources nécessaires aux systèmes à base de connaissances.

Nos travaux sur ce thème s'inscrivent dans un courant de recherche actif en France depuis une dizaine d'années, qui est en partie animé par le groupe de travail « Terminologie et Intelligence Artificielle » dont Brigitte Biébow, Adeline Nazarenko et Sylvie Szulman sont membres. Ces travaux rejoignent les recherches plus récentes qui sont menées au niveau européen dans la perspective de la construction d'un web sémantique. L'ancienneté de nos recherches et la coopération étroite entre le traitement automatique des langues et l'ingénierie des connaissances, donne une place originale à l'équipe RCLN dans ce paysage scientifique.

Nous mettons ici l'accent sur la constitution des bases de connaissances terminologiques et conceptuelles (ontologies).

- **Construire des ressources terminologiques à partir de corpus**

La construction des ressources lexicales constitue souvent le goulot d'étranglement des systèmes de traitement de corpus : les dictionnaires de langue sont généralement peu adaptés, les ressources lexicales spécialisées sont rarement disponibles et toujours difficiles à réutiliser. Nous développons des méthodes et des outils d'aide à la construction de ces ressources terminologiques pour un corpus et en fonction d'une tâche donnée. Nous travaillons dans deux directions spécifiques : l'extraction des termes et celle des relations qu'ils entretiennent.

Extraction de termes Sophie Aubin et Thierry Hamon ont proposé une méthode hybride qui combine l'acquisition de connaissances en corpus et l'exploitation de ressources terminologiques existantes. Les résultats montrent que la combinaison des deux approches

permet d'obtenir de meilleurs résultats que l'une ou l'autre des deux approches prise isolément. Cette méthode a été implémentée dans l'extracteur de termes YaTeA (pour Yet Another Term ExtrActor) [CI-22]. Le système effectue une analyse syntaxique superficielle d'un corpus pour identifier des termes candidats et utilise en parallèle une ressource terminologique du domaine pour adapter le processus d'extraction de termes à ce domaine en reprenant en compte certaines de ses caractéristiques. L'exploitation de ressources lexicales ne va cependant pas de soi : il est important d'identifier leur adéquation au corpus. Adeline Nazarenko et Goritsa Ninova ont proposé une définition formelle de la couverture à partir de quatre métriques qui donnent une vue globale de la pertinence d'une ressource pour un corpus [CO-28].

Extraction de relations terminologiques Nos premiers travaux avaient porté sur le repérage de liens de synonymie à l'aide d'un dictionnaire de langue générale, sur le calcul de similarité sémantique, et sur l'extraction de relation d'hyperonymie à l'aide de patrons lexico-syntaxiques [RE-8]. Touria Aït El Mekki et Adeline Nazarenko ont poursuivi dans cette voie en montrant comment ces différentes méthodes peuvent collaborer pour construire un réseau sémantique de termes. On peut ainsi construire des index de documents structurés où les différentes entrées de l'index sont organisées en niveaux hiérarchiques [CI-11].

• Construire des ontologies à partir de corpus

Les méthodes de construction d'ontologies restent un thème de recherche important de l'équipe RCLN : il s'agit d'utiliser les textes commme sources de connaissances. Brigitte Biébow et Sylvie Szulman ont développé Terminae, une méthode et un outil d'aide à la construction d'ontologies à partir de textes, qui ont été présentés en 1999. Cette méthode s'appuie sur une analyse préalable des corpus qui repose sur les principes linguistiques et des logiciels de TAL présentés ci-dessus.

La méthode Terminae continue d'évoluer, d'une part dans la recherche d'éléments linguistiques permettant de modéliser les connaissances ontologiques, d'autre part dans l'intégration de ressources existante. B. Biébow et S. Szulman ont présenté dans [CL-7] un guide pratique de modélisation pour construire des ressources ontologiques à partir de textes. Le guide expose en détail comment l'utilisation de logiciels de traitement automatique des langues et de modélisation peut aider un analyste lors du dépouillement et de la fouille dans les données textuelles. Il précise pas à pas les différentes étapes du travail en mettant en évidence la navigation entre les nombreux éléments de modélisation. La méthode devrait prochainement intégrer les relations linguistiques trouvées dans le corpus et utiles à la description d'un domaine [?].

L'outil Terminae a également évolué. Il permet désormais la visualisation des résultats des analyseurs syntaxiques tels Syntex[5] et Yatea []. Il intègre les résultats de Synoterm[6]. Le langage de description d'ontologie de Terminae autorise l'importation et l'exportation en OWL [CO-11]. La technologie de plugin a été étudiée et une première implantation a été créée.

Ces différents travaux ont été effectuées dans le cadre de projets collaboratifs (ATONANT, LOIS) ou partenariat bilatéraux (TENNAXIA).

• Réutiliser des ressources termino-ontologiques existantes

Une autre évolution de la méthode Terminae devrait permettre d'intégrer des ontologies génériques préexistantes pour soulager le travail de modélisation. En collaboration avec Sylvie Després (CRIPS 5 - équipe IAD), Sylvie Szulman étudie différents aspects du problème : l'alignement des ontologies et l'intégration d'ontologies de haut-niveau comme (Dolce) [?]. La figure ci-après présente la méthode Terminae résultante qui a été testée essentiellement dans le domaine juridique, plus particulièrement l'étude de directives européennes [?].

Lorsque les ontologies produites sont utilisées pour l'analyse ou l'annotation de textes, au-delà de la fusion des ontologies acquises à partir de corpus et des ontologies préexistantes, se pose la question de l'articulation des connaissances qui servent à l'interprétation. Elles sont de différents niveaux : connaissances terminologiques et ontologiques ressources morphologiques et

[5]Développé par D. Bourigault, ERSS.
[6]Développé par T. Hamon.

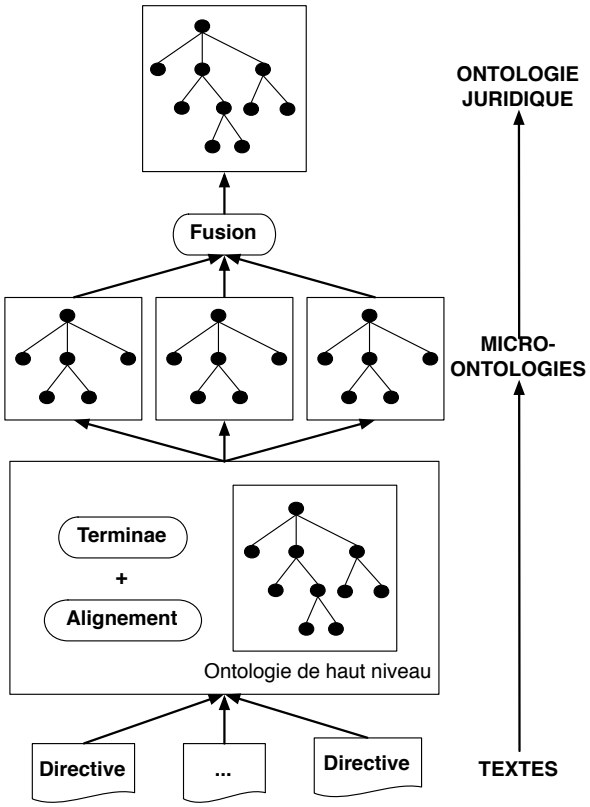


FIG. 1.4: Méthode de construction d'ontologie : application au domaine juridique.

syntaxiques utilisées pour l'analyse ou règles de désambiguïsation et d'extraction mobilisées dans les tâches d'étiquetage et d'extraction. L'articulation de ces différents niveaux de connaissances représente un défi important pour tout ce qui relève de l'annotation sémantique de documents et partant, du web sémantique. Cette question, que nous avons abordée de manière éparse à travers différents projets, constitue le sujet central de la thèse de Frederik Cailliau. Dans un contexte industriel, celui-ci développe un système complet de gestion des ressources linguistiques exploitées par le moteur de recherche de Sinequa, Intuition [CO-41]. Cette réflexion de fond, essentielle pour développer de véritables systèmes de compréhension de textes devrait être prolongée dans les prochaines années.

1.4 Bilan et perspectives de recherche

Cette section dresse le bilan des recherches menées au sein de l'équipe RCLN sur la période 2004-2007 et montre dans quelles directions ces travaux vont se prolonger dans les prochaines années.

1.4.1 Une équipe renforcée

La période considérée a permis à l'équipe RCLN de renforcer ses compétences et son potentiel de recherche.

- **Les moyens d'une recherche expérimentale**

L'équipe RCLN a conduit des recherches assez fondamentales en TAL depuis longtemps. La fin des années 1990 et le début des années 2000 ont été l'occasion d'ouvrir un nouvel axe autour de l'analyse de corpus qui comporte une part significative de travail expérimental. Un important effort de développement a donc été réalisé au cours de la période écoulée pour se doter d'un ensemble d'outils de traitement et de manipulation de corpus textuels ainsi qu'une plate-forme qui les intègre [CI-19]. Cet investissement facilite aujourd'hui la réalisation de prototypes de recherche, il permet de faire des expériences d'envergure sur corpus (analyse sémantique, construction d'ontologie) et de mieux valoriser nos travaux de recherche sous la forme d'outils opérationnels (système IndDoc, par ex. [RE-12]). Cet effort doit être poursuivi dans les prochaines années pour maintenir et faire évoluer les outils existants en fonction de nos besoins de recherche. L'équipe RCLN souhaite recruter un ingénieur de recherche permanent pour assurer la pérennité de ces développements et coordonner le travail des ingénieurs temporaires recrutés sur différents contrats.

- **Un ensemble élargi de compétences du TAL à l'apprentissage**

Depuis quelques années, le LIPN a mis en avant un axe inter-équipes pour le développement de travaux combinant les techniques du TAL et de l'apprentissage. Cette collaboration s'est mise en place sur la période considérée, grâce au recrutement de maîtres de conférences avec des compétences mixtes (2 dans l'équipe RCLN, 1 dans l'équipe A^3) et dans le cadre de différents projets (apprentissage de règles d'extraction, par ex.). Depuis 2006, les deux équipes RCLN et A^3 sont engagées dans un projet commun (Infom@gic) pour le développement d'outils de traitement de l'information textuelle.

- **Une forte implication dans les projets collaboratifs**

Composée de 12 membres permanents auxquels viennent s'ajouter des doctorants, des chercheurs post-doctorants, des ingénieurs de recherche et des stagiaires, l'équipe RCLN a désormais atteint une taille critique. Sur la période considérée, elle a participé (parfois comme coordinateur) à de nombreux projets qui portaient aussi bien sur des axes de recherche fondamentaux que sur des thématiques plus appliquées. Cette politique doit se poursuivre dans les prochaines années : certains projets ont été lancés en 2006 (projet RNTL TextCoop autour de l'analyse de la typologie des textes, projet ANR MICRAC sur l'analyse du raisonnement causal, projet Infom@gic pour

la fouille de données textuelles), d'autres en 2007 (projet ANR RNTL DAFOe de développement d'une plateforme de construction d'ontologies) et d'autres encore doivent démarrer en 2008 (projet Quaero pour le développement de moteurs de recherche multimedia, projet ANR Masse de données CROTAL). La diversité de ces projets et, pour certains, leur envergure montrent que l'équipe RCLN est aujourd'hui reconnue sur l'ensemble de ses axes de recherche, aussi bien au plan national qu'au plan européen et international.

1.4.2 Un dialogue renouvelé

Si les années antérieures avaient permis à un nouvel axe de voir le jour, la période 2004-2007 a permis de renforcer le dialogue interne de l'équipe.

- **Un projet fédérateur concernant l'accès à l'information médicale**

Le lancement du projet ACTMed a permis à l'ensemble de l'équipe RCLN de collaborer sur un même projet en lien avec le LLI et le LIM&Bio. Ce travail a mis en évidence des synergies entre les différentes directions de recherche de l'équipe RCLN : les outils de recherche d'information et de navigation que nous avons proposés exploitent des techniques de TAL et d'ingénierie des connaissances mais ils reposent aussi sur une analyse fine de la sémantique des textes et du raisonnement médical. Ce projet a également fait émerger de nouvelles thématiques : le problème de la structuration des documents est ainsi apparu comme central dans la plupart de nos travaux sur les Guides de Bonne Pratique ; la question de la complémentarité des diverses techniques classiques d'accès à l'information (recherche et extraction d'information, question-réponse, navigation) est également apparue comme importante, l'objectif étant désormais de permettre à l'utilisateur de passer dynamiquement et naturellement d'une activité de recherche documentaire à la consultation de résumés ou à la navigation dans le document, voire à l'interrogation de documents [CL-5], dans une forme de « compréhension à profondeur variable ».

- **L'émergence de préoccupations transversales**

L'évolution des travaux de l'équipe a fait apparaître des questions communes aux différents thèmes de recherche.

A la suite des travaux sur le contenu des documents (analyse de la terminologie, des entités nommées, de la causalité, etc.), un travail récent a porté sur des structures discursives plus larges, comme les séquences conditions-actions au sein de documents procéduraux (voir section 1.3.2). Des besoins récurrents sont apparus autour de ce thème encore récent au sein de l'équipe. Cet axe de recherche va donc être approfondi. Ces travaux reposent sur une analyse précise des textes et devraient permettent à terme d'améliorer l'accès au document sous forme électronique.

Cet axe de recherche implique également l'identification de genres textuels, afin d'obtenir des ensembles de textes homogènes sur lesquels les outils d'analyse linguistique peuvent s'appliquer de façon pertinente. Cette problématique est étudiée au sein du projet TextCoop. Pour mener à bien ces recherches, il est apparu nécessaire de faire appel à la fois à une analyse fine des marqueurs textuels et à des méthodes plus robustes d'analyse de corpus.

Nos recherches sur la sémantique des verbes se situent également à la croisée de nos différents thèmes de recherche puisqu'il s'agit aujourd'hui d'apprendre des schémas prédicatifs à partir de corpus pour améliorer l'extraction d'information [CI-3] (voir ci-dessous).

1.4.3 De nouveaux défis

- **La compréhension fine, des questions anciennes toujours d'actualité**

Nos recherches antérieures en sémantique lexicale concernaient notamment le traitement de la polysémie des verbes. Elles s'attachaient à mettre au jour les influences croisées entre le verbe et ses arguments souvent nécessaires pour faire émerger l'interprétation de chaque composant. Les projets actuels de l'équipe RCLN amènent à relancer ces recherches dans une perspective plus appliquée qu'initialement : on a besoin de prendre en compte la sémantique

des verbes pour mettre en relation des entités nommées qui auraient été extraites des textes (extraction d'information) ou pour établir des relations ontologiques « transversales » (ingénierie des connaissances textuelles). Au-delà des indices de surface, nous proposons de prendre en compte une analyse sémantique locale qui repose sur l'acquisition et la reconnaissance de schémas prédicatifs en corpus. Une thèse est en cours sur ce sujet (Cédric Messiant).

Différentes recherches ont été menées au cours de la période autour de la problématique du raisonnement et de sa modélisation. Sur les différentes thèmes considérés (calcul des événements, identification des normes qui structurent le raisonnement, analyse du raisonnement médical), nous avons montré la puissance explicative de nos modèles et proposé des pistes pour articuler ces modélisations avec l'analyse linguistique. Ces travaux sont bien reconnus au plan international. Ils se poursuivent dans le cadre du projet MICRAC et avec la thèse de Sara Boutouhami.

- **L'ingénierie des connaissances, un rendez-vous à ne pas manquer**

L'équipe RCLN a investi depuis une dizaine d'années dans la problématique de la construction d'ontologies à partir de textes. Ces travaux sont aujourd'hui reconnus au niveau européen. Au-delà de la seule construction d'ontologies, nos recherches s'étendent désormais à l'ensemble des opérations de l'ingénierie ontologique (fusion, alignement d'ontologies) dont l'enjeu devient central avec le développement du web sémantique et des outils de recherche multimedia qui présupposent une annotation sémantique des documents. En dépit de la faiblesse numérique de cette thématique au LIPN, les collaborations extérieures ont permis de maintenir une activité importante (participation au projet européen LOIS, , projet RNTL ATONANT, partenariat industriel). Cette activité se développe encore à la faveur de projets où le LIPN apporte son expertise dans la construction de ressources termino-ontologiques (projet RNTL DAFOE et projet Quaero, prévu pour 2008). Les principaux enjeux concernent désormais l'intégration de ressources existantes et la coopération de méthodes hétérogènes dans la construction d'ontologies.

- **La sémantique de corpus à la rencontre de vrais enjeux**

L'émergence de moteurs de recherche sémantiques distribués lance de nouveaux défis au TAL. Dans la lignée du projet ALVIS et en lien avec les recherches actuelles sur les architectures distribuées de moteurs de recherche, nous cherchons à définir des modèles sémantiques permettant à la fois de caractériser un document au sein d'une base documentaire et un moteur de recherche dans son réseau de pairs, de manière à permettre un traitement intelligent de la requête de l'utilisateur qui ne sait pas nécessairement à quel moteur de recherche s'adresser. Ce projet fait appel aux techniques d'acquisition de ressources termino-ontologiques de l'équipe RCLN et ouvre de nouvelles pistes de travail sur la catégorisation de documents et l'alignement d'ontologies. Une étude préliminaire est en cours entre le LIPN (Antoine Rozenknop et Adeline Nazarenko) et l'EPFL (Martin Rajman) mais ce projet a vocation à être élargi à d'autres partenaires, si les premières pistes de travail se révèlent fructueuses.

La maturité de nos travaux d'analyse de corpus nous invite également à repenser nos méthodes d'évaluation, l'objectif étant de mieux mesurer l'avancée du TAL et leur adéquation aux besoins applicatifs. Sur la période considérée, nous avons commencé à nous intéresser plus directement à ces problèmes d'évaluation que posent les techniques de TAL et d'ingénierie des connaissances textuelles [CL-1]. Cette réflexion a été initiée par notre participation à quelques campagnes d'évaluation et par l'organisation d'ateliers. Une thèse sur ce thème a été lancée en coopération avec ELDA. Cet effort va s'intensifier dans les années à venir. Une réflexion de fond a été initiée dans le cadre du groupe de travail TIA et l'équipe RCLN doit jouer un rôle actif dans l'évaluation au sein du projet Quaero sur certaines tâches de TAL (élaboration des protocoles d'évaluation et organisation de campagnes).

1.5 Synthèse

Très soudée autour des problématiques de la représentation des connaissances et du traitement automatique des langues, l'équipe RCLN présente une orientation scientifique originale dans le paysage français. Même si elle a encore besoin d'être consolidée par le recrutement d'un maître

de conférence ou d'un chargé de recherche en ingénierie des connaissances et par le recrutement d'un ingénieur de recherche permanent, l'évolution de l'équipe RCLN sur les dernières années (recrutement de nouveaux maîtres de conférences en TAL, financement sur contrat de personnel temporaire) lui a permis d'atteindre une taille critique : elle mène de front des recherches théoriques, des recherches plus appliquées et des travaux de développement quand nécessaire. Cette équipe jouit désormais d'une bonne reconnaissance au niveau local, national et européen, comme en témoignent, la diversité de ses collaborations, le nombre de ses publications et la participation à des projets d'envergure.

1.6 Animation de la recherche

1.6.1 Evaluation de la recherche

- **Comités de rédaction**

  - *Cognitive Science Quarterly*, 2000-2004 ; D. Kayser
  - *Intellectica*, 1996-2004 ; D. Kayser
  - *Traitement Automatique des Langues*, numéros spéciaux, 2006, 2007 ; A.Nazarenko, T.Poibeau

  Daniel Kayser fait en outre partie du comité éditorial d'une collection co-édité par Ophrys et la Maison des Sciences de l'Homme.

- **Comités de programmes**

  - EuroCogSci 07, *European Conference on Cognitive Science*, mai 2007, Delphes, Grèce ; D. Kayser (co-président)
  - RFIA, *Reconnaissance des Formes et Intelligence Artificielle*, 2004, 2006 ; D. Kayser (comité restreint)
  - TALN *Traitement Automatique des Langues Naturelles*, 2005 ; D. Kayser
  - IWCS *International Workshop on Computational Semantics*, 2005, 2007 ; D. Kayser
  - SARA *Symposium on Abstraction, Reformulation and Approximation*, 2005, Edinburgh, UK ; D. Kayser
  - ISPS *International Symposium on Programming and Systems*, 2005 ; D. Kayser
  - TIA *Terminologie et Intelligence Artificielle*, 2005, 2006, 2007 ; D. Kayser, A. Nazarenko, S. Szulman
  - RECITAL *Rencontre de Etudiants-Chercheurs en Traitement Automatique des Langues*, 2004, 2005, 2006, 2007 ; T. Hamon
  - Texte et Connaissances, colloque organisé dans le cadre de la semaine de la Connaissance, 2006 ; S. Després, S. Szulman
  - SemAnnot *International Workshop on Knowledge Markup and Semantic Annotation*, 2004-2006 ; S. Szulman
  - EKAW *Knowledge Engineering and Knowledge Management Managing Knowledge*, 2004-2005 ; B. Biébow
  - IC *Ingénierie des Connaissances*, 2004-2007 ; B. Biébow, S. Després, S. Szulman
  - OntoTexte *Atelier scientifique "Ontologies et Textes"*, 2007 ; A. Nazarenko
  - OLP *ECAI Workshop on Ontology Learning/Population*, 2004 ; A. Nazarenko
  - SMBM *2nd International Symposium on Semantic Mining in Biomedicine*, 2006 ; A. Nazarenko
  - PSB *New frontiers in biomedical text mining, PSB workshop*, 2007 ; A. Nazarenko
  - MEMURA *LREC workshop on Methodologies and Evaluation of Multiword Units in Real-world Applications*, 2004 ; A. Nazarenko
  - LREC *Language Resources and Evaluation Conference*, 2006 ; A. Nazarenko
  - KDLL *PKDD workshop on Knowledge Discovery in Life Science Literature*, 2006 ; A. Nazarenko
  - CIFT, *Conférence Internationale sur la fouille de textes*2004 ; A. Nazarenko
  - FinTAL *5th International Conference on Natural Language Processing*, 2006 ; A. Nazarenko

- RJCRI 2007, *Rencontres Jeunes Chercheurs en Recherche d'Information*, 2007 ; H. Zargayouna
- SAAKM *Semantic Authoring, Annotation and Knowledge Markup Workshop*, 2007 ; S. Szulman
- KW4ED Workshop *Knowledge Management and Semantic Web for Engineering Design Workshop*, 2007 ; S. Szulman
- RFIA *Reconnaissance des Formes et Intelligence Artificielle* 2008 ; D. Kayser (comité restreint)
- ISPS *International Symposium on Programming and Systems* 2007 ; D. Kayser
- SARA *Symposium on Abstraction, Reformulation and Approximation* 2007 ; D. Kayser
- KRAQ*Knowledge and Reasoning for Answering Questions* 2005 (satellite de IJCAI), 2006 (satellite de EACL) ; D. Kayser
- EGC *Extraction et Gestion des Connaissances*, 2005 et 2008 ; S. Després, S. Szulman
- Toth *Conférence Terminologie et Ontologie : Théorie et Applications*, 2007-2008 ; S. Després
- Atelier RAPC *Atelier Raisonnement à partir de cas*, 2005-2008 ; S. Després
- OTJ *Journée Ontologies et Textes Juridiques, organisée dans le cadre de la semaine de la connaissance*, 2006 ; S. Després, S. Szulman

• Comités de lecture

Les membres de l'équipe assurent régulièrement des relectures d'articles pour des conférences et des revues parmi lesquelles

- *International Joint Conference on Artificial Intelligence* (conférence)
- *European Conference on Artificial Intelligence* (conférence)
- *Data & Knowledge Engineering* (revue)
- *Human Language Technology* (revue)
- *Techniques et Sciences Informatiques* (revue)
- *Revue des Nouvelles Technologies de l'Information* (revue)
- *Traitement Automatique des langues* (revue)
- *Linguisticae Investigationes* (revue)
- *Information - Interaction - Intelligence* (revue)

• Comités d'évaluation

- (membre) Jury d'attribution des prix de thèses européennes; Daniel Kayser, comme ECCAI Fellow.
- (membre) Comité national du CNRS, section 45 ; 2003 et 2004 ; Daniel Kayser.
- (président) Jury d'attribution du prix de thèse de l'ASTI, 2006 ; Daniel Kayser.
- (membre) Jury d'attribution du prix de thèse de l'ASTI, 2007 ; Daniel Kayser.
- (président) Jury d'admissibilité des directeurs de recherche (CNRS, section 45, 2003-04) ; Daniel Kayser .
- (président) Comité d'évaluation du LORIA (Univ. Nancy 1 et 2, CNRS & INRIA), 2004 ; Daniel Kayser.
- (président) Comité scientifique disciplinaire n°1 de l'ANR, 2007 ; Daniel Kayser
- (membre) Comité scientifique du Pôle "Intelligence Logicielle" ; Daniel Kayser (contrat Etat-Région Lorraine).
- (membre) Comité scientifique du Crip5 (Univ. Paris 5) entre 2001 et 2005 ; Daniel Kayser.
- (membre) Commission technique "Ingénierie des Connaissances" du Pôle de compétitivité CapDigital depuis 2005 ; Adeline Nazarenko
- (membre) Divers comités d'évaluation de laboratoires et de formations pour le MENESR ; Daniel Kayser.
- (membre et président en 2006) Jury d'attribution de Primes d'encadrement doctoral depuis plusieurs années; Daniel Kayser.
- (expert) Comité de préparation du 7ème PCRD, 2006 ; Daniel Kayser .

- **Expertises de projets**

  - ANR Programme *Blanc*, 2005, 2006 (D. Kayser), 2007 (A. Nazarenko)
  - ANR Programmes *RNTL* et *Technologies logicielles*, 2005, 2006, 2007 (A. Nazarenko)
  - ANR Programme *Masse de données*, 2006, 2007 (A. Nazarenko, T. Poibeau)
  - ANR Programme *Jeunes Chercheurs*, 2007 (A. Nazarenko)
  - ANR Programme *Corpus*, 2007 (A. Nazarenko, F. Gayral)
  - ANR Programme *TECSAN*, 2007 (A. Nazarenko)
  - Projets-contrats de la région de la Basse-Normandie (A. Nazarenko)
  - Projets de recherche canadien, 2005-2006 (A. Nazarenko)
  - Projets de recherche de la Swiss National Science Foundation, 2006 (A. Nazarenko)

- **Jurys**

  - Jurys de thèse (hors Université Paris 13) : D. Kayser (3 dont 2 comme rapporteur), F. Lévy (2 comme rapporteur), A. Nazarenko (10 dont 7 comme rapporteur et 1 à l'étranger), S. Szulman (2)
  - Jury d'Habilitation à diriger des Recherches (hors Université Paris 13) : D. Kayser (2 comme rapporteur)
  - Jurys de thèse (hors Université Paris 13) : S. Després (2 dont 1 comme rapporteur)

1.6.2 Organisation de la recherche

- **Organisation d'événements scientifiques et conférences**

Les membres de l'équipe RCLN ont organisé ou participé à l'organisation de divers événements scientifiques et conférences.

  - Participation à l'organisation du *Genic Interaction Extraction Challenge* associé au workshop LLL-ICML 2005, 6 équipes officiellement participantes
  - Co-organisation de l'atelier *Evaluation des méthodes de terminologie computationnelle*, associé à la conférence "Terminologie et Intelligence Artificielle (TIA)", Rouen, 2005, 50 participants.
  - Co-organisation de l' *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* associé à la "20th International Conference on Computational Linguistics (COLING)", 2004, Geneve, 70 participants.
  - Organisation de l'*International Workshop on Terminology design : quality criteria and evaluation methods (TermEval)*, may 2006, in conjunction with the *Language Resources and Evaluation Conference (LREC)*, 2006, Gènes, Italie, 30 participants.

- **Organisation de séminaires**

L'équipe RCLN organise depuis 2001 des séminaires « Connaissances, Langue et Informatique » à raison de trois séminaires par an en moyenne. Cette formule de séminaire « en duo » fait intervenir deux orateurs sur une demi journée et permet de confronter deux visions différentes sur un même sujet. L'objectif est que ces présentations donnent lieu à un dialogue non seulement entre les orateurs, mais également avec l'assistance. Le séminaire est destiné aussi bien aux chercheurs confirmés qu'aux étudiants s'initiant au domaine abordé. Ce séminaire est largement ouvert à l'extérieur. Selon les thèmes abordés, l'audience varie entre 20 et 50 participants.

L'équipe RCLN accueille également régulièrement des chercheurs qui viennent présenter leurs travaux. Ces séminaires, à visée plus interne, contribue à la vie scientifique de l'équipe.

1.6.3 Contrats

Type	Coordinateur	Resp. scient.	Nom	Financeur	Période	Durée (mois)	Budget global (Keuros)	Part LIPN (Keuros)
Bioinfo.		A. Nazarenko	Caderige	inter-EPST	11/02-10/03	12		50
RNTL	•	A. Nazarenko	ExtraPloDocs	MENESR	12/02-09/05	33	390	170
RNTL		S. Szulman	ATONANT	MENESR	09/02-11/05	38	671	31
RNTL		T. Hamon	CESART	MENESR	01/04-01/06	36	142	8
CMEP-CNRS		D. Kayser	Modèles de l'action	EGIDE	01/02-12/05	48	80	15
TCAN-CNRS	•	F. Lévy	OGRE	CNRS	07/03-12/05	30	40	10
STREP		A. Nazarenko	ALVIS	UE	10/04-12/06	36	2 258	210
PPF	•	A. Nazarenko	ACTmed	MENESR	10/06-12/08	48	200	85
ANR		D. Kayser	MICRAC	ANR	01/06-12/08	36	325	64
RNTL		F. Gayral	TextCoop	ANR	01/06-12/08	36	424	128
Cap Digital		A. Nazarenko	Infom@gic	DGE	12/05-11/06	12		48
RNTL		S. Szulman	DAFOe	ANR	01/07-12/09	36	800	95

Caderige *Autres partenaires : AÏDA-IRISA, Leibniz-IMAG, LRI, INRA-MIG, INRA-ENSAR.*

Caderige - *Catégorisation Automatique de Documents pour l'Extraction de Réseaux d'Interactions GEniques* - Ce projet visait au développement d'une application d'interrogation en langage naturel de la base Medline pour les biologistes, l'objectif étant dans un premier temps d'exploiter les notices bibliographiques portant sur la transcription des gènes. Une méthode fondée sur l'extraction d'information et l'acquisition de connaissances a été définie [CO-19]. Cette première étude s'est poursuivie dans le projet ExtraPloDocs (http://caderige.imag.fr/).

ExtraPloDocs *Autres partenaires : MIG (INRA), Isoft*

ExtraPloDocs - *EXTRAction d'information pour l'exPLOration de la DOCumentation Scientifique* - Dans le cadre de ce projet, ont été réalisés des outils permettant de 1) extraire des informations en génomique à partir de grandes bases textuelles comme Medline ; 2) acquérir, à partir d'un échantillon de corpus, les ressources nécessaires à l'analyse des textes (dictionnaires d'entités nommées, terminologies, ontologies) ; 3) apprendre les règles d'extraction par des méthodes d'apprentissage relationnel. Le LIPN qui a assuré la coordination du projet et a principalement contribué aux deux premières tâches (http://www-lipn.univ-paris13.fr/RCLN/projets/ExtraPloDocs/).

ATONANT *Autres partenaires : INSA, LaRIA, LIP6 (Univ. Paris 6 & CNRS), EADS, CEA*

ATONANT - *ATONANT* (RNTL-Technolangue) - L'objectif du projet était de développer des outils d'aide à l'enrichissement semi-automatique d'ontologies destinés à une plate-forme modulaire. Le domaine d'application était la médecine, plus précisement l'enrichissement de ressources du CISMeF (Catalogue et Index des Sites Médicaux Francophones). Le LIPN a participé au couplage de l'outil Terminae avec un éditeur mettant en oeuvre la méthode OntoSpec développée au LaRIA. Nous avons aussi adapté Terminae pour traiter les données fournies par le CHU de Rouen et créé dans le langage de description de Terminae une ontologie relative à la

virologie. L'importation et l'exportation en OWL d'ontologies ont été implémentés afin de comuniquer nos résultats à nos partenaires dans un langage standard (http://atonant.insa-rouen.fr/).

CESART *Autres partenaires : ELDA, Cersates (Univ. Lille 3), EDF, Temis, RALI, CEA*

CESART - *Campagne d'Evaluation de Systèmes d'Acquisition de Ressources terminologiques* (RNTL-Technolangue) - Ce projet a permis de mener une campagne d'évaluation des outils d'acquisition de ressources terminologiques à partir de corpus écrits (http://www.technolangue.net/article.php3?id_article=200/). Le LIPN est intervenu dans la définition du protocole d'évaluation et comme fournisseur de logiciel évalué, SynoTerm, pour l'acquisition de relations sémantiques (synonymie) entre termes. La campagne d'évaluation s'est déroulée en juin 2005 sur deux corpus du domaine de la médecine (pages web de Santé Canada) et de l'éducation (numéros de la revue SPIRALE).

Modèles de l'action *Autres partenaires : Université des Sciences et Techniques Houari Boumediène, Alger*

Modèles de l'action - Ce contrat, conclu initialement pour prolonger le travail sur les modèles de l'action commencé dans les thèses d'Université (française) et d'État (algérienne) d'Aïcha Mokhtari sous la direction de Daniel Kayser a permis à la partie algérienne à reprendre des activités de recherche en liaison avec l'équipe RCLN mais aussi avec d'autres équipes du LIPN (des thèses seront bientôt soutenues sous la direction de Laure Petrucci, équipe LCR). De son côté, la partie française a repris ses travaux antérieurs sur la causalité avec notamment l'apport de Sara Boutouhami (étudiante de magister d'Aïcha Mokhtari, qui va commencer sa thèse dans l'équipe RCLN).

OGRE *Autres partenaires : Limsi (CNRS), GREYC (Univ. Caen & CNRS), Dyalang (Univ. Rouen & CNRS)*

OGRE - *Intervalles temporels et applications à la linguistique textuelle* - Ce projet a réuni des linguistes et des informaticiens et a jeté les bases d'une analyse et d'une représentation des phrases itératives (ex : « le lundi, il mange à la cantine »). Avec ses partenaires, le LIPN a montré la nécessité de prendre en compte types d'intervalles pour représenter des événements itératifs selon que l'accent est mis sur la continuité du procès global, sur son éclatement ou sur certains de ses composants. Deux formalismes ont été étudiés pour représenter ces différentes valeurs aspectuelles. A des fins d'expérimentation, le LIPN a constitué un corpus de 18000 phrases itératives annotées extraites du journal *Le Monde*. (http://www-lipn.univ-paris13.fr/~levy/TCAN/).

ALVIS *Autres partenaires : Helsinki Institute for Information Technology, Unité Mathématique, Informatique et Génome (INRA), Ecole Polytechnique Fédérale de Lausanne, Lund University, Technical University of Denmark, Index Data Aps, Exalead SA, ALMA Bioinformatica S.L., Jozef Stefan Institute, Tsinghua University*

ALVIS - *Superpeer Semantic Search Engine* - Ce projet vise au développement d'un moteur de recherche sémantique et distribué. Les aspects sémantiques assurent une meilleure adéquation des résultats aux requêtes formulées par les utilisateurs. La recherche distribuée repose sur la coopération de différents serveurs spécialisés dans des domaines différents. Le LIPN est responsable de l'enrichissement sémantique des documents et a réalisé la plateforme d'annotation de corpus pour l'anglais et le français. Le LIPN participe aussi à la tâche d'asquisition des ressources spécialisées qui, pour un domaine donné, sont utilisées par le module d'annotation : ont été développés un extracteur de termes, une méthode de spécialisation pour un parseur, un système d'apprentissage de règles de résolution des anaphores (http://www.alvis.info).

ACTMED *Autres partenaires : Lim&Bio (Univ. Paris 13), LLI (Univ. Paris 13 & CNRS*

ACTmed - *Accès au Contenu des Textes Médicaux : Analyse des besoins et développement d'outils* - Ce projet réunit trois laboratoires de l'Université Paris-Nord qui collaborent pour développer des méthodes d'accès à l'information textuelle médicale. Il s'agit 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). En complément de l'expertise des autres laboratoires dans l'ingénierie des systèmes de santé et la

linguistique, le LIPN analyse la structure des documents et le raisonnement médical sous-jacent et développe des prototypes intelligents d'interrogation et de consultation des guides.

MICRAC *Autres partenaires : IRIT (Univ. Toulouse 3 & CNRS), laboratoire Dynamiques Sociocognitives et Vie Politique (Univ. Toulouse 2), Centre de Recherche en Informatique de Lens (Univ. Artois & CNRS)*

MICRAC - *Modèles Informatique et Cognitifs du raisonnement Causal* - Ce projet pluridisciplinaire (http://www.irit.fr/MICRAC/), qui repose sur une étroite collaboration entre chercheurs en intelligence artificielle et en psychologie cognitive, a pour ambition la construction d'un système informatique capable de représenter des relations causales, de reconnaître ces relations dans un ensemble d'événements ordonnés, de fournir des explications quant au déroulement de ces événements, et d'attribuer des responsabilités à des agents parties prenantes de ces événements. Dans ce projet, l'équipe RCLN s'intéresse aux constats d'accident. Elle cherche à faire valider par des protocoles appropriés les conjectures effectuées jusqu'ici sur ce qu'un lecteur standard considère comme cause de l'accident, à comparer différentes théories formelles du raisonnement causal pour voir celles qui correspondent le mieux aux résultats observés, et enfin à implémenter un système de raisonnement causal opérant selon ces principes (thèse de Farid Nouioua à soutenir début 2007, et thèse de Sara Boutouhami commençant en octobre 2006).

TextCoop *Autres partenaires : IRIT(Univ Toulouse 3 & CNRS, SINEQUA*

TextCoop - *Typologies et grammaires de textes afin de rendre un TEXTe plus «COOPératif» en gestion de l'information* - Le projet TextCoop a 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). Dans ce projet, le rôle du LIPN concerne l'analyse des types de texte (manuel utilisateur, recommandations médicales, documentation technique, inventaire, article scientifique, discours, etc.) et la conception des méthodes qui permettant de prendre en compte le type des textes 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.

Infom@gic *Autres partenaires : une vingtaine de partenaires industriels (Thales, EADS, ONERA, Bertin, Vecsys, Sinequa...) et institutionels (Universités Paris 6, Paris 8, Marne La vallée, GET, Limsi-CNRS.*

Infom@gic - Ce projet s'inscrit dans le cadre du pôle de compétitivité francilien Cap Digital - IMVN (Image, multimédia et vie numérique) qui est consacré aux technologies de l'information et de la communication. Infom@gic vise 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 plate-forme commune qui doit couvrir les grands domaines de l'analyse d'information quelles que soient les sources (données structurées, texte, images et sons) : la recherche et l'indexation, l'extraction de connaissances et la fusion d'informations multimédias. Le LIPN apporte à ce projet son expertise à la fois en apprentissage artificiel (équipe A^3) et en fouille de données textuelles (équipe RCLN).

DAFOe *Autres partenaires : INSERM UMR_S 872, ENST GET, IRIT, LIPN, LISI, Mondeca, Supelec, UTC.*

DAFOe - *Differential and Formal Ontologies Editor For Aapplications* L'objectif de ce projet est de proposer une méthode complète associée à une plateforme technique pour concevoir des ontologies, de la modélisation à partir du domaine à leur évolution en passant par leur formalisation et exploitation. S'appuyant sur les acquis de travaux antérieurs, à la fois issus des partenaires et de la littérature du domaine, le projet a pour but de prendre en charge la modélisation sémantique des concepts ontologiques (différentiels et formels) pour motiver et justifier les représentations formelles qui seront utilisées et en faciliter la révision. La plateforme technique DAFOe est un ensemble d'outils dont un éditeur d'ontologies qui prend en charge toute la question de la sémantique de ces ontologies, à travers des questions épistémologiques liées aux concepts

formels de haut niveau ? *i.e.* la top ontology ou ontologie formelle ? et, vis-à-vis de la composante métier, à travers des travaux sur les corpus textuels. Le but est ainsi de rendre disponible, à travers la plateforme, les méthodologies qui vont permettre de construire l'ontologie. On obtient ainsi une ontologie formalisée qui pourra être traitée dans un éditeur dontologie respectant les standards des langages d'ontologies du W3C (OWL).

1.6.4 Dissémination

- **Distinctions**

A. Nazarenko s'est vu décerner la médaille de bronze du CNRS (section 45, 2004).

- **Logiciels**

Terminae est une plate-forme d'aide à la construction de terminologies et d'ontologies à partir de l'étude de textes. Il apporte une aide à la modélisation d'un domaine à partir de textes. Il intègre un concordancier et les résultats d'outils de traitement de corpus[7]. La version actuelle permet l'importation d'une ontologie en RDFS et OWL et l'exportation d'une ontologie Terminae en OIL et OWL.

Distribution : contacter Sylvie Szulman.

IndDoc est un outil d'aide à la construction d'index de fin de livre. Il suit une méthodologie bien définie qui exploite l'apport de la terminologie computationnelle et intégre des outils liguistiques[8]. L'objectif de IndDoc est de mettre à la disposition des indexeurs (documentalistes, éditeurs, auteurs de livre, etc.) une nomenclature riche ainsi qu'une liste de renvois. Les indexeurs peuvent valider ou valider les entrées de l'index produit ainsi que spécifiifier un paramétrage (contraintes éditoriales, etc.) pour la mise en forme de l'index final (sous forme papier ou numérique).

Une présentation détaillée de l'outil ainsi qu'une démo sont disponibles sur le site Web IndDoc : http ://www-lipn.univ-paris13.fr/RCLN/outils/IndDoc/

TagEN est un logiciel de reconnaissance d'entités nommées à base d'automates à nombre fini d'états (reposant sur le logiciel Unitex de l'Université de Marne-la-Vallée). Des ressources (lexiques et grammaires) existent pour le français et l'anglais et des systèmes noyaux pour d'autres langues (ressources à étendre pour les rendre utilisables). Le logiciel est open source et fonctionne essentiellement sous Linux (les ressources sont utilisables sous Windows, linux et Mac). Voir http ://www-lipn.univ-paris13.fr/ poibeau/tagen.html.

Distribution : contacter Thierry Poibeau.

YaTeA (Yet another Term extrActor) 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 têtes et modifieurs. 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 (désambiguisation exogène) [CI-22]. Ces termes attestés aident au repérage des groupes, à leur analyse syntaxique et participent également à l'extraction des termes candidats. YaTeA sera prochainement disponible sous licence GPL sous la forme d'un module CPAN. D'autres informations sont disponibles à l'adresse suivante : http ://www-lipn.univ-paris13.fr/ aubin/yatea.html

BiolG est une version modifiée de l'analyseur syntaxique Link Grammar Parser de Sleator et Temperley, adapté au domaine biomédical telle que décrite dans [CI-13]. BioLG présente deux extensions par rapport à la version originale de Link Grammar : il intègre un système de spécification de règles de "morpho-guessing" par expressions régulières et prend en compte un pré-étiquetage morpho-syntaxique du corpus à analyser.

[7] Lexter et/ou Syntex développés par D. Bourigault (didier.bourigault@univ-tlse2.fr), SynoTerm et Yatea développés au LIPN par T. Hamon et S. Aubin.

[8] Tels que YaTeA développé par S. Aubin et T. Hamon ou Faster développé par C. Jacquemin

BioLG est accessible et téléchargeable sous licence GPL à cette adresse : http://www.it.utu.fi/biolg/. BiolG a été réalisé en collaboration avec Sampo Pyysalo, Département de Technologie de l'Information de l'université de Turku, Finlande.

Ogmios La platforme d'annotation linguistique développé dans le cadre du projet ALVIS, offre la possibilité d'enrichir linguistiquement des documents textuels, rédigés en français ou en anglais. Ce processus d'annotation est à la fois robuste et rapide, du fait le l'architecture modulaire et distribuée.

La platforme propose plusieurs types de traitements linguistiques : reconnaissance d'entités nommées, segmentation en mots et en phrases, lemmatisation, étiquetage morpho-syntaxique, étiquetage terminologique, et analyse syntaxique (uniquement sur l'anglais, pour l'instant).

La plateforme d'annotation linguistique est développée en Perl. Elle est disponible en open-source, sous forme de module CPAN (http://search.cpan.org/~thhamon/Alvis-NLPPlatform-0.1/), ainsi que sous paquetage Debian[9].

SynoTerm SynoTerm est un logiciel de détection de relations de synonymie entre termes. Il peut être utilisé sur le français ou l'anglais, soit pour l'indexation libre d'un corpus de textes (calcul des relations de synonymie entre les termes extraits d'un corpus), soit pour l'indexation contrôlée d'un corpus de textes : il identifie les relations de synonymie entre des termes extraits d'un corpus et ceux issus d'une liste de termes contrôlés (terminologie, thesaurus).

L'interface de SynoTerm permet d'assister le travail de constitution de terminologie structurée : les relations proposées par SynoTerm doivent être validées par un utilisateur.

SynoTerm est développé en C et en Perl (utilisation de Perl/Tk pour l'interface graphique). Il sera bientôt disponible en open-source et sa mise à disposition sur SourceForge (http ://sourceforge.net/) est à l'étude.

http://www-lipn.univ-paris13.fr/~hamon/SynoTerm/SynoTerm.html

Concordanta Concordanta est un concordancier qui permet d'effectuer des recherches dans des fichiers préalablement enrichis d'annotations linguistiques. Cet outil repose sur la librairie LOX (C++). Il fonctionne sur les systèmes les plus courants (Linux/Macintosh/Windows).

Disponibilité : contacter L. Audibert.

XExplora et Explora Ces outils permettent de rechercher, d'observer des phénomènes linguistiques dans des corpus écrits et de formater librement le résultat des recherches. Ces applications exploitent la librairie C++, LOX, qui implémente le concept de méta-expressions régulières et son application sur des corpus. Ces logiciels fonctionnent sur les systèmes les plus courants (Linux/Macintosh/Windows).

Disponibilité : contacter L. Audibert.

• Actions de vulgarisation

- T. Aït El Mekki : Fête de la science/Savantes banlieues, Université Paris 13, 2004 :démonstration ("Construction automatique d'index de fin de livres")
- A. Nazarenko : Fête de la science/Savantes banlieues, Université Paris 13, 2004, 2006, 2007 : mini-conférences ("La fouille de texte ou comment trouver une aiguille dans une botte de textes")
- A. Nazarenko : participation à un déjeuner de la technologie organisé par organisé par différents acteurs de la Seine Saint Denis sur le thème "Convergences technologiques - GESTIOgestion de documents électroniques : défis et innovation", 23 mars 2007 "

• Invitations, tutoriels, école d'été

- D. Kayser : conférence invitée au '6th International Symposium on Programming and Systems (ISPS)', 2003

[9] http://www-lipn.univ-paris13.fr/~hamon/ALVIS/Debian/testing/libalvis-nlpplatform-perl_0.1-1_all.deb

- T. Poibeau : conférence invitée au 'Workshop of the network of excellence on text mining and statistics (Nemis)', 2004
- D. Kayser : conférence invitée au '1st Indian Conference on Logic and Applications', 2005
- D. Kayser : conférence invitée à la journée scientifique 'Langage, origines et développement', 2005
- T. Poibeau : co-organisation de la table-ronde « Future trends in Question Answering » avec M. Minock, dans le cadre du workshop « knowledge and Reasoning for Question Answering » (Worlshop IJCAI KRAQ), 2005
- D. Kayser : participation à la table ronde 'Langage et Cognition' lors de l'atelier de préparation du 7ème PCRD, 2006
- A. Nazarenko : participation à la table ronde 'TAL, document et compréhension' animée par B. Victorri dans le cadre l'école de printemps 'Langue, connaissances, information', 2005
- A. Nazarenko : participation à la table ronde 'Evaluation of terminological tools and products' dans le cadre du 'Workshop lors de la journée 'Texte et connaissances' de la semaine de la connaissance, 2006
- A. Nazarenko et S. Szulman : animation et participation à la table ronde 'Marqueurs linguistiques et extraction de connaissances à partir de textes' lors de la journée 'Texte et connaissances' de la semaine de la connaissance, 2006

- **Sociétés savantes**

Les membres de l'équipe RCLN sont adhérents de différentes sociétés savantes. Ils ont joué un rôle actif dans l'ATALA (Association pour le Traitement Automatique des LAngues) et l'ARCo (Association pour la Recherche Cognitive).

1.6.5 Collaborations et Invitations

- **Collaborations internationales (hors contrats de recherche)**

- Coopération franco-algérienne : cette coopération a permis le séjour au LIPN pour des durées d'une à six semaines et des séjours d'une semaine à Alger pour 4 chercheurs du LIPN. Plusieurs publications ont été effectuées dans ce cadre [] et deux thèses sont en cours d'achèvement.
- Partenariat avec les Universités roumaines (Politehnica de Bucarest, Univ. Stefan Cel Mare de Suceava) dans le cadre de l'Université Virtuelle Francophone : ce partenariat a permis la réalisation d'un cours sur le web, la co-rédaction d'un livre [], la participation à des écoles d'été et diverses missions bilatérales.
- Protocole de coopération scientifique dans le cadre du projet européen eContent LOIS (Lexical Ontologies for legal Information Sharing), avec l'Italian National Research Council (IT), les Universités de Vienne (AT), Sheffield, Evora (PT), Leiden (NL), l'Académie des Sciences de la République Tchèque (CS) (http ://www.loisproject.org/modules/addresses), 2005-2006.
- Participation au projet STIC-Asie (coordonné par P. Saint-Dizier) : T. Poibeau a visité le NAIST (Nara, Japon) en mai 2006 et présenté les recherches du LIPN en matière de Traitement des Langues ; il donnera une série de cours lors de l'école d'été Language, Artificial Intelligence and Computer Science for Natural Language Processing applications (LAICS-NLP, Bangkok, 16-22 oct. 2006) sur les thèmes suivants.

- **Collaborations universitaires (hors contrats de recherche)**

- Coopération avec l'INSERM (U729) : participation conjointe à une campagne d'évaluation des systèmes de question-réponse (T. Poibeau)
- Participation à l'action spécifique RIP-WEB (recherche d'informations précises sur le web) coordonnée par le Limsi-CNRS (T. Poibeau)
- Participation au GDR 2521 "Sémantique et Modélisation" (C. Recanati)

- Participation au groupe de travail "Terminologie et Intelligence artificielle (TIA, AFIA, GDR I3) (B. Biébow, A. Nazarenko, S. Szulman)
- Membre du bureau du réseau thématique pluridisciplinaire « Langage naturel, logique et philosophie du langage » (section 34 et 35 du CNRS, 2005-2006) (A. Nazarenko)
- Participation au Groupe de travail Léonard (Acquisition du langage et grammaticalisation, A. morgenstern) : conseil pour l'informatisation des corpus – programme Childes et base de données Chat (T. Poibeau)

- **Collaborations industrielles (hors contrats de recherche)**

  - Contrat d'encadrement de master recherche avec la société TENNAXIA et en collaboration avec le LaRIA (Univ. Amiens), 2006 (S. Szulman)
  - Convention Cifre avec la société Sinequa (2004-2007, thèse de F. Cailliau)
  - Convention Cifre avec la société ELDA (2005-2008, thèse de O. Hamon)
  - Collaboration avec Isoft : expertise du LIPN pour l'intégration d'outils linguistiques au sein de la plate-forme de traitement de données Amadea (T. Poibeau)
  - Collaboration avec Exalead et encadrement d'un stagiaire de Master Recherche pour l'enrichissement automatique de thesaurus à partir de données issues du web (T. Poibeau)

- **Autres échanges universitaires**

  - Séjour d'un mois de T. Poibeau au Japon à l'Université de Tokyo et au National Institute of Informatics, oct-nov. 2004
  - Séjour de trois semaines de Sampo Pyysalo (Université de Turku, FI) au LIPN, juin 2006 : collaboration autour du développement d'un analyseur syntaxique adapté à l'analyse des textes de biologie [CI-13]
  - Séjour de 3 semaines en juin 2006 de Sophia Ananiadou (U. de Manchester, NaCTeM National Center on Text Mining) au LIPN comme professeure invitée

1.7 Formation à la recherche

1.7.1 Habilitations à Diriger les Recherches

Brigitte Biébow HDR en informatique, Université Paris 13, « De Dasert à Terminae : voyage au pays de la modélisation des connaissances à partir de textes », soutenue le 9 juillet 2004
Coordinateur : D. Kayser
Situation actuelle du diplômé : retraite

Adeline Nazarenko HDR en informatique, Université Paris 13, « Donner accès au contenu des documents textuels Acquisition de connaissances et analyse de corpus spécialisés », soutenue le 14 décembre 2004
Coordinateur : D. Kayser
Situation actuelle du diplômé : Professeur à l'Université Paris 13

1.7.2 Thèses soutenues

Touria Aït El Mekki Doctorat en informatique, Université Paris 13, « Construction semi-automatique d'index de fin de livre », soutenue le 21 décembre 2004
Directeurs : D. Kayser et A. Nazarenko
Financement : allocataire-monitrice
Situation actuelle du diplômé : Maître de Conférences à l'Université d'Angers

Laure Bourgois Doctorat en informatique, Université Paris 13, « Représentation et comparaison de Web services complexes avec des logiques dynamiques », soutenue le 19 juin 2007 avec mention Très Honorable.
Directeur : F. Levy
Financement : 3 ans de contrat France-Télécom R&D, un an Ater
Situation actuelle de la diplômée : enseignante en école d'ingénieur à Dakkar.

Valentina Ceausu Doctorat en informatique, Université Paris 5, « Définition d'un cadre sémantique pour la catégorisation de données textuelles. Application à l'accidentologie », soutenue le 18 juin 2007
Directeurs : S. Despres
Financement : allocataire-monitrice - ATER
Situation actuelle du diplômé : ATER à l'IUT de l'Université Paris 13

Farid Nouioua Doctorat en informatique, Université Paris 13, « Extraction et Utilisation des Normes pour un Raisonnement Causal dans un Corpus Textuel », soutenue le 25 Avril 2007 avec la mention très honorable avec félicitations
Directeur : D. Kayser
Financement : allocataire-moniteur
Situation actuelle du diplômé : un post-doctorat à l'Université Paris 11-Orsay.

1.7.3 Thèses en cours

Rim Bentebibel inscrite en thèse d'informatique à l'Université Paris 5, depuis 10/2004
Sujet : « Adaptation dans le contexte du raisonnement à partir de cas textuel »
Directeur : S. Despres
Financement : allocataire-moniteur - ATER à l'université Paris 5

Aurélien Bossard inscrit en thèse d'informatique à l'Université Paris 13, depuis 10/2006
Sujet : « Evolution et fusion d'informations extraites de corpus en langage naturel »
Directeurs : D.Kayser et T. Poibeau
Financement : allocataire-moniteur

Amanda Bouffier inscrit en thèse d'informatique à l'Université Paris 13, depuis 11/2004
Sujet : « Structuration de textes pour l'aide à la modélisation »
Directeurs : D. Kayser et T. Poibeau
Financement : allocataire-moniteur
Situation actuelle : ATER

Sara Boutouhami-Nouioua inscrit en thèse d'informatique à l'Université Paris 13, depuis fin 2006
Sujet : « Etude de l'aspect argumentatif dans les textes d'accidents de la route »
Directeur : D. Kayser
Financement : projet ANR

Frederik Cailliau inscrit en thèse d'informatique à l'Université Paris 13, depuis 11/2004
Sujet : « Conception et mise en oeuvre d'un système de gestion de ressources linguistiques »
Directrice : A. Nazarenko
Financement : Bourse Cifre (Sinequa)

Olivier Hamon inscrit en thèse d'informatique à l'Université Paris 13, depuis 11/2005
Sujet : « Méthodes et les protocoles l'évaluation des technologies de traitement automatique des langues et de la communication Homme-Machine »
Directrice : A. Nazarenko
Financement : Bourse Cifre (ELDA)

Cédric Messiant inscrit en thèse d'informatique à l'Université Paris 13, depuis 10/2006
Sujet : « Acquisition automatique d'informations linguistiques sur corpus »
Directeurs : A. Nazarenko et T. Poibeau
Financement : bourse DGA et monitorat

Sonia Mhiri inscrit en thèse d'informatique à l'Université Paris 5, co-tutelle avec la faculté des sciences de Tunis, depuis 10/2005
Sujet : « Intégration d'ontologies »
Directeurs : S. Despres & E. Zagrouba
Financement : bourse de l'état tunisien

Thibault Mondary inscrit en thèse d'informatique à l'Université Paris 13, depuis 10/2006
Sujet : « Construction d'ontologies à partir de textes »
Directeurs : S. Després et A. Nazarenko
Financement : allocataire-moniteur

Davy Weissenbacher inscrit en thèse d'informatique à l'Université Paris 13, depuis 10/2003
Sujet : « Incomplétude et Imperfection des informations pour l'Extraction d'Information, un cadre applicatif : la résolution de l'anaphore pronominale »
Directeurs : C. Fouqueré et A. Nazarenko
Financement : allocataire-moniteur
Situation actuelle : ATER

1.7.4 Encadrement de stages de recherche

- 2003 (A. Nazarenko) : Z. Qian (DEA IA-OC, P13&P8), *Etude de la contribution de la terminologie au filtrage de fragments de textes*
- 2004 (A. Nazarenko) : G. Ninova (Master 2 recherche, P7), *Mesures de l'adéquation des ressources lexicales et ontologiques en corpus*
- 2004 (A. Nazarenko) : F. Cailliau (DEA IA-OC, P13&P8), *Vers une gestion intégrée des ressources linguistiques*
- 2004 (T. Poibeau) : J.-F. Berroyer (DEA IA-OC, P13&P8), *TagEN : un système de repérage des entités nommées*
- 2005 (S. Salotti) : I. Mansour (Master 2 recherche, P9), *Etiquetage sémantique et désambiguïsation d'un corpus médical*
- 2005 (S. Szulman) : F.-E. Calvier (Master 2 recherche, P13), *Alignement et fusion d'ontologies*
- 2005 (T. Poibeau) : A.-M. Renoux (Master 2 professionel, INaLCO), *Apport de la textométrie pour l'analyse de textes littéraires*
- 2006 (D. Kayser) : A. Gheerbrant (Master 2 recherche, P13), *Vers une formalisation non monotone des guides de bonne pratique médicale*
- 2006 (H. Zargayouna et S. Salotti) : V. Rosas (Master 2 professionnel, Univ. Aix-Marseille 2, *Analyse et structuration de requêtes en langage naturel pour l'interrogation de documents médicaux*
- 2006 (A. Bouffier et A. Nazarenko) : T. Mondary (Master 2 recherche, P13), *Etude de la navigation dans un document structuré. Application aux guides de bonne pratique médicale*
- 2006 (T. Poibeau) : A. Bossard (Master 2 recherche, P13), *Utilisation et acquisition de schémas prédicatifs pour l'extraction d'information*
- 2006 (C. Recanati et Y. Bennani) : N. Rogovschi (Master 2 recherche, P13), *Systèmes d'apprentissage non supervisé connexionnistes et stochastiques pour la fouille de données structurées en séquences*
- 2006 (T. Hamon) : J. Derivière (Master 2 recherche, P13), *Etude de l'impact du traitement automatique des langues en recherche d'information*
- 2006 (T. Poibeau) : T. Moal (Master 2 recherche, P13), *Utilisation de listes et de structures de pages web pour l'enrichissement de thesaurus*
- 2007 (D. Kayser) : N. Kamenoff (Master 2 recherche, P13), *Analyse sémantique, Modélisation et Visualisation :application aux constats d'accidents de la route*
- 2007 (S. Després et S. Szulman) : M. Drine (Master 2 recherche, P13), *Construction d'une ressource termino-ontologique à partir de propositions de scénarios d'aménagements urbains*
- 2007 (D. Weissenbacher et A. Nazarenko) : A.A. Bouchouareb (Master 2 recherche, P13), *Fouille de textes pour l'enrichissement d'une base de données de noms de gènes*
- 2007 (H. Zargayouna et A. Nazarenko) : M. Rakho (Master 1, P7), *Génération et visualisation d'index de fin de livre*

1.7.5 Cours de DEA ou de Master 2 recherche

A. Nazarenko est responsable du Master 2 Recherche en Informatique de l'Université Paris 13 (spécialité « Modélisation Informatique des Connaissances et du raisonnement »).

- B. Biébow, D. Kayser, F. Levy, A. Nazarenko, S. Szulman : cours de DEA IA-OC (P13) : *Ingénierie des connaissances textuelles*, 2003-2005.

- B. Biébow, S. Szulman : cours de DEA 127 (P9) : *Ontologies et Web sémantique*, 2003-2004.
- D. Kayser : cours de Master 2 recherche MICR (P13) : *Traitement de l'incertain et de la causalité*, 2005-07.
- F. Lévy : cours de Master 2 recherche MICR (P13) : *Formalismes de représentation des connaissances*, 2005-06.
- F. Gayral, D. Kayser, A. Nazarenko, A. Rozenknop : cours de Master 2 recherche MICR (P13) : *Traitement automatique des langues* [CL-8], 2005-06.
- T. Hamon, A. Nazarenko, S. Szulman : cours de Master 2 recherche MICR (P13) : *Ingénierie des connaissances textuelles*, 2005-07.
- R. Besançon, A. Nazarenko, T. Poibeau : cours de Master 2 recherche MICR (P13) : *Recherche et extraction d'information*, 2005-06.
- A. Nazarenko : cours de Master 2 recherche MICR (P13) *Méthodologie de la recherche*, 2005-06.
- S. Szulman : cours de DEA 127 (P9) : *Ontologies et Web sémantique*, 2004-2005.
- S. Szulman : cours de Master 2 (P9) : *Ingénierie des ontologies et Terminae*, 2005-2006.
- H. Zargayouna et A. Nazarenko : cours de Master 2 pro MÉTIERS DU LIVRE (P13) *Traitement du document numérique*, 2006, 2007.

1.8 Publications principales

Livres

[POIBEAU :2003 :CCSD-00005506 :1] Thierry Poibeau. *Extraction automatique d'information : Du texte brut au Web sémantique*. Lavoisier, 2003, ISBN 2-7462-0610-2.

Chapitres de livres

[NAZARENKO :2006 :CCSD-00097092 :1] Adeline Nazarenko. *Le point sur l'état actuel des connaissances en Traitement Automatique des Langues*, Gérard Sabah, éditeur, *Compréhension des langues et interaction*, pages 31–70. Hermès, 2006.

[BIEBOW :2005 :CCSD-00089085 :1] Brigitte Biébow, Sylvie Szulman et Nathalie Aussenac-Gilles. *Modélisation du domaine par une méthode fondée sur l'analyse de corpus*, R. Teulier ; J. Charlet ; P. Tchounikine, éditeur, *Ingénierie des connaissances*, pages 49–71. L'Harmattan, 2005.

[GAYRAL :2005 :CCSD-00084948 :1] Françoise Gayral, Daniel Kayser et François Lévy. *Challenging the Principle of Compositionality in Interpreting Natural Language Texts*, E. Machery, M. Werning et G. Schurz, éditeurs, *The Compositionality of Meaning and Content*, volume 2, pages 83–105. Ontos Verlag, 2005.

[NAZARENKO :2005 :CCSD-00098028 :1] Adeline Nazarenko. *Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ?*, Anne Condamines, éditeur, *Sémantique et corpus*, pages 211–244. Traité IC2, série Cognition et traitement de l'information. Lavoisier, 2005.

[NEDELLEC :2005 :CCSD-00097093 :1] Claire Nédellec et Adeline Nazarenko. *Ontology and Information Extraction : a necessary symbiosis*, Bernardo Magnini Paul Buitelaar, Philipp Cimiano, éditeur, *Ontology Design and Population*, pages 155–170. IOS Press, 2005.

[POIBEAU :2005 :CCSD-00005507 :1] Thierry Poibeau. *Content annotation for the semantic Web*, Spiros Sirmakessis, éditeur, *Knowledge Mining*, pages 250–262. Springer, 2005. ISBN 3-540-25070-0.

[POIBEAU :2004 :CCSD-00085173 :1] Adeline Nazarenko et Thierry Poibeau. *L'évaluation des systèmes d'analyse et de compréhension de textes*, Stéphane Chaudiron, éditeur, *L'évaluation des systèmes de traitement de l'information*, pages 127–148. Hermès, 2004. ISBN 2-7462-0862-8.

Articles dans des revues internationales avec comité de lecture sélectif

[KAYSER :2005 :CCSD-00091643 :1] Daniel Kayser et Farid Nouioua. About norms and causes. *International Journal on Artificial Intelligence Tools*, 14 :7–23, 2005.

[NAZARENKO :2005 :CCSD-00098030 :1] Adeline Nazarenko et Touria Aït El Mekki. Building back-of-the-book indexes. *Terminology*, 11(1) :193–218, 2005.

[KAYSER :2003 :CCSD-00091608 :1] Daniel Kayser. Abstraction and natural language semantics. *Philosophical Transactions of the Royal Society of London*, 358 :1261–1268, 2003.

Articles dans des revues nationales ou internationales

2006 [RE-6*] Davy Weissenbacher. Les réseaux bayésiens : un formalisme adapté au traitement automatique des langues ? *Revue d'Intelligence Artificielle, Numéro spécial : "Modèles Graphiques Probabilistes"*, 2006. À paraître.

[ZARGAYOUNA :2006 :CCSD-00092307 :1] Haifa Zargayouna, Touria Ait El Mekki, Laurent Audibert et Adeline Nazarenko. IndDoc : An aid for the back-of-the-book indexer. *The Indexer*, 25(2) :122–125, 2006.

[RECANATI :2005 :CCSD-00085004 :1] Catherine Recanati. Raisonner avec des diagrammes : perspectives cognitives et computationnelles. *Intellectica*, 40 :9–42, 2005.

[AITELMEKKI :2004 :CCSD-00098041 :1] Touria Aït El Mekki et Adeline Nazarenko. L'index de fin de livre, une forme de résumé indicatif ? *Traitement Automatique des Langues*, 45(1) :121–150, 2004.

[GRABAR :2004 :CCSD-00084817 :1] Natalia Grabar et Thierry Hamon. Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'Intelligence Artificielle*, 18/1 :57–85, 2004.

[LEVY :2004 :CCSD-00088705 :1] François Lévy. Statistiques et contrefactuels, quelques explications. *Intellectica*, 38 :187–191, 2004. Introduction à "Expérimentation, inférence statistique et analyse causale" de B. Lecoutre dans le même numéro.

Communications dans des conférences internationales avec comité de lecture

[AUBIN :2006 :CCSD-00090074 :1] Sophie Aubin et Thierry Hamon. Improving Term Extraction with terminological resources. Tapio Salakoski, Filip Ginter, Sampo Pyysalo et Tapio Pahikkala, éditeurs, *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, LNAI 4139, pages 380–387. Springer, 2006.

[DERIVIERE :2006 :CCSD-00090079 :1] Julien Derivière, Thierry Hamon et Adeline Nazarenko. A scalable and distributed NLP architecture for Web document annotation. Tapio Salakoski, Filip Ginter, Sampo Pyysalo et Tapio Pahikkala, éditeurs, *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, LNAI 4139, pages 56–67. Springer, 2006.

[PYYSALO :2006 :CCSD-00082533 :1] Sampo Pyysalo, Tapio Salakoski, Sophie Aubin et Adeline Nazarenko. Lexical adaptation of Link Grammar to the biomedical sublanguage : a comparative evaluation of three approaches. *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006)*, pages 60–67. Sophia Ananiadou, Juliane Fluck, 2006.

[WEISSENBACHER :2006 :CCSD-00095189 :1] Davy Weissenbacher. Bayesian Network, a model for NLP ? *European Conference of the Association for Computational Linguistics (EACL 2006)*, pages 195–198, 2006.

[AITELMEKKI :2005 :CCSD-00098036 :1] Touria Aït El Mekki et Adeline Nazarenko. Using NLP to build the hypertextuel network of a back-of-the-book index. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 316–320. Institute for Parallel Processing, Bulgarian Academy of Sciences, 2005.

[AUBIN :2005 :CCSD-00082542 :1] Sophie Aubin, Adeline Nazarenko et Claire Nédellec. Adapting a general parser to a sublanguage. 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'05)*, pages 89–93. INCOMA Ltd., 2005.

[KAYSER :2005 :CCSD-00091645 :1] Daniel Kayser. What logic can and cannot do in natural language semantics. *First Indian Conference on Logic and its relationship with other disciplines*, Actes en cours d'édition, Bombay, Inde, 2005.

[GAYRAL :2004 :CCSD-00092218 :1] Françoise Gayral, Daniel Kayser et François Lévy. Challenging the Principle of Compositionality in Interpreting Natural Language Texts. *Conference on Compositionality, Concepts and Cognition*, pages 111–127, Düsseldorf, Allemagne, 2004.

[KAYSER :2004 :CCSD-00091633 :1] Daniel Kayser et Farid Nouioua. Representing knowledge about norms. *Proc. 16th European Conference on Artificial Intelligence*, Valencia, Spain, pages 363–367. IOS Press, 2004.

[POIBEAU :2004 :CCSD-00009168 :1] Thierry Poibeau. Automatic extraction of paraphrastic phrases from medium size corpora. Sergei Nirenburg, éditeur, *Actes de la conférence Computational Linguistics (COLING'2004)*, pages 638–644. Association for Computational Linguistics, 2004.

Workshops et autres conférences avec comité de lecture

[BOUFFIER :2006 :CCSD-00085170 :1] Amanda Bouffier et Thierry Poibeau. Segmentation et structuration de textes procéduraux pour l'aide à la modélisation de connaissances : le rôle de la structure visuelle. *Semaine de la Connaissance*, pages 195–205, Nantes, France, 2006.

[CAILLIAU :2006 :CCSD-00098149 :1] Frederik Cailliau. Unifier la gestion de ressources linguistiques en contexte multilingue. *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2006*, pages 455–461. Presses universitaires de Louvain, 2006.

[NAZARENKO :2006 :CCSD-00094950 :1] Adeline Nazarenko, Claire Nédellec, Erik Alphonse, Sophie Aubin, Thierry Hamon et Alain-Pierre Manine. Semantic annotation in the ALVIS project. *Proceeding of IIIA-2006 : International Workshop on Intelligent Information Access*, Helsinki, Finlande, 2006.

[NINOVA :2005 :CCSD-00084841 :1] Goritsa Ninova, Adeline Nazarenko, Thierry Hamon et Sylvie Szulman. Comment mesurer la couverture d'une ressource terminologique pour un corpus ? *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2005*, pages 293–302. ATALA, 2005.

[ALPHONSE :2004 :CCSD-00098040 :1] Erick Alphonse, Sophie Aubin, Philippe Bessières, Gilles Bisson, Thierry Hamon, Sandrine Lagarrigue, Adeline Nazarenko, Alain-Pierre Manine, Claire Nédellec, Mohamed Ould Abdel Vetah, Thierry Poibeau et Davy Weissenbacher. Event-based information extraction for the biomedical domain : the Caderige project. Nigel Collier, Patrick Ruch et Adeline Nazarenko, éditeurs, *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (COLING'04)*, pages 43–49, Geneva, Suisse, 2004.

[GAYRAL :2004 :CCSD-00092223 :1] Françoise Gayral, Daniel Kayser et François Lévy. The role of time in considering collections. *Journées de Sémantique et Modélisation*, pages 30–34, Lyon, France, 2004.

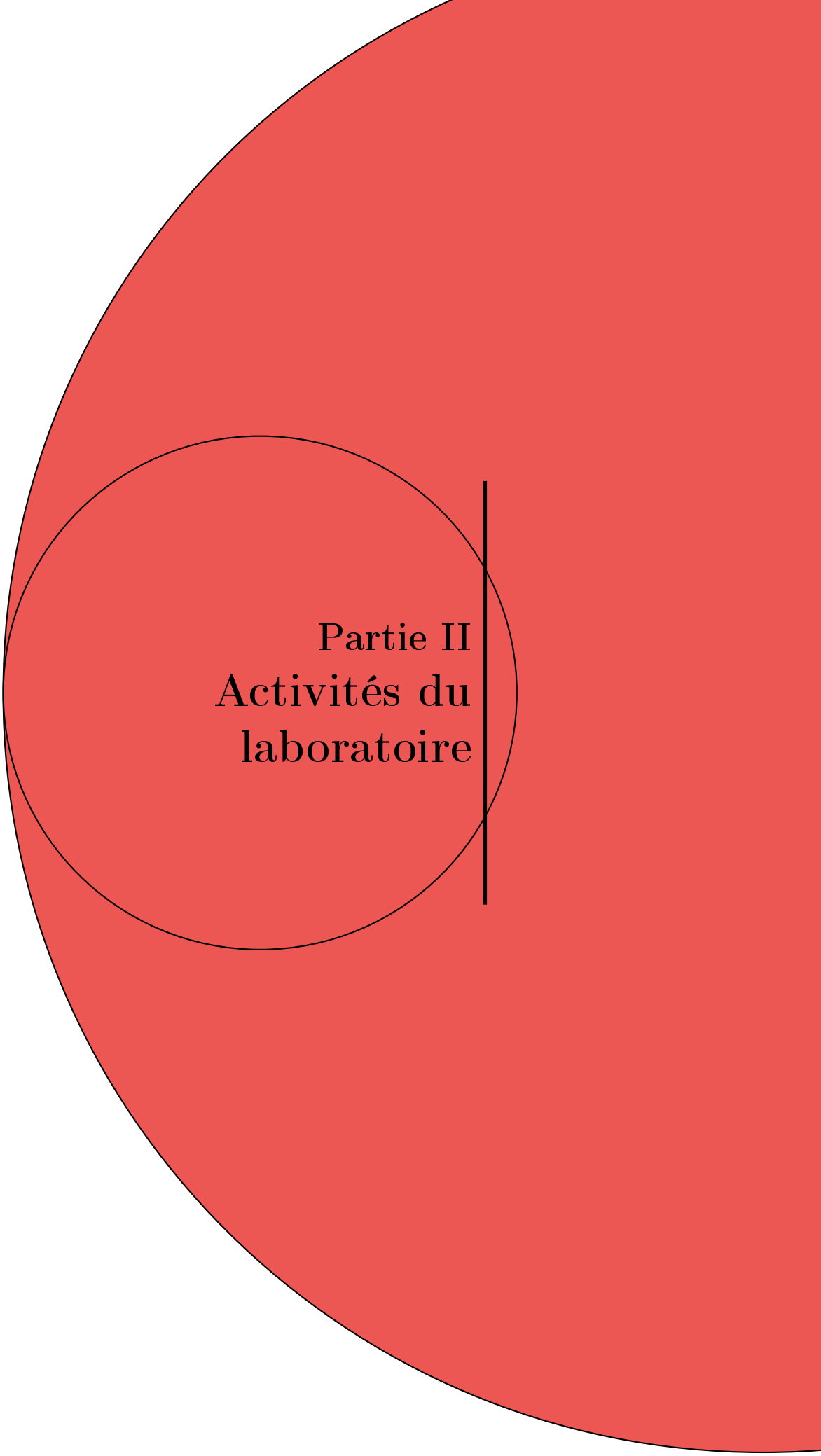
[SZULMAN :2004 :CCSD-00087557 :1] Sylvie Szulman et Brigitte Biébow. OWL et Terminae. *IC'2004 - 15ème journées francophones d'Ingénierie des Connaissances*, pages 41–52. Presses Universitaires de Grenoble, 2004.

[ZARGAYOUNA :2004 :CCSD-00094619 :1] Haifa Zargayouna et Sylvie Salotti. Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. *15ème journées francophones d'Ingénierie des Connaissances (IC'2004)*, pages 249–260. Presses Universitaires de Grenoble, 2004.

[GAYRAL :2003 :CCSD-00092208 :1] Françoise Gayral, Daniel Kayser et François Lévy. Collections évolutives et partitions. *Colloque "La partition en langue et en discours"*, France, 2003.

Outils

[DERIVIERE :2006 :CCSD-00090753 :1] Julien Derivière et Thierry Hamon. Alvis NLP platform. A scalable architecture using existing NLP tools to annotate large collections of web documents, 2006.



Partie II
Activités du laboratoire

Chapitre 1
Formation par la recherche

1.1 Master Informatique

Le Master (Sciences, Technologie, Santé) **mention Informatique** offre deux spécialités de formation qui s'appuient notamment sur des axes forts et des compétences bien développées au LIPN :

une spécialité à finalité professionnelle : Exploration Informatique des Données (EID)

une spécialité à finalité recherche : Modélisation Informatique des Connaissances et du Raisonnement (MICR)

La spécialité EID a pour objet l'acquisition de connaissances approfondies dans des domaines particuliers complémentaires : l'informatique et l'extraction de connaissances à partir de données. Cette spécialité est fondée sur l'acquisition de techniques destinées à favoriser l'exercice d'un métier déterminé dans le domaine de l'exploration des données.

La spécialité MICR a pour objectif de fournir aux étudiants les éléments fondamentaux leur permettant d'entamer une recherche tant en intelligence artificielle qu'en traitement automatique du langage, en spécifications de systèmes complexes ou en théorie de la programmation. Dans cette spécialité, plusieurs parcours-types sont possibles, on peut cependant en préfigurer quelques uns :

– Parcours « Traitement Automatique des Langues Naturelles » : Traitement automatique

des langues, Ingénierie des connaissances textuelles, Recherche et extraction d 'information, Ressources lexicales ;
- Parcours « Spécifications et Programmation » : Systèmes parallèles asynchrones, Systèmes infinis, temporisés et hybrides, Programmes et preuves, Unité d'approfondissement au choix ;
- Parcours « Logique et Programmation » : Calculs de processus, Programmes et preuves, Concurrence, Unité d'approfondissement au choix ;
- Parcours « Apprentissage Artificiel et Fouille de Données » : Apprentissage numérique, Apprentissage symbolique et RaPC, Apprentissage statistique, Fouille de données relationnelles.

Formation par la recherche

Ce master est ouvert à la pluridisciplinarité qui est un des atouts de l'université Paris 13, avec par exemple une UE de linguistique sur les ressources lexicales et une UE sur la bio-informatique.

1.2 Master Mathématique Informatique

Le Master (Sciences, Technologie, Santé) **mention Mathématique et Informatique** a pour objectif de fournir aux étudiants les éléments fondamentaux leur permettant d'acquérir une véritable double compétence en Informatique et en Mathématiques et leurs applications. Elle offre l'accès à des cours très complémentaires dans les domaines de l'informatique, des mathématiques discrètes et continues, des phénomènes aléatoires et de la combinatoire. Elle permet, outre les aspects théoriques, de toucher aux problèmes applicatifs sources de modèles complexes. La formation s'appuie à la fois sur les laboratoire LIPN et LAGA de l'université Paris 13.

Outre les thèses possibles dans les laboratoires, en particulier au LAGA ou au LIPN, les étudiants peuvent se tourner vers les activités de recherche et développement en entreprise. Cette seconde voie n'est pas à négliger, les entreprises devenant un endroit où les compétences solides en mathématiques et informatique appliquées, attestées par une formation par la recherche, sont de plus en plus recherchées. Le master mathématiques et informatique vise à donner au travers de la formation le bagage indispensable d'un bon Ingénieur en Recherche et Développement ou d'un Consultant auprès d'entreprises de divers secteurs de pointe.

La première année est consacrée aux apprentissages des outils de base en mathématiques appliquées et en informatique. En fin d'année, un travail personnel d'étude et de recherche mené en coordination avec l'équipe pédagogique doit permettre à l'étudiant de mettre par lui-même en application les connaissances acquises pendant l'année. La seconde année est l'occasion d'approfondissement des connaissances en choisissant parmi de nombreux sujets : résolution numérique d'équations aux dérivées partielles, recherche opérationnelle, combinatoire, mathématiques financières.

1.3 Thèses et Habilitations

Thèses en cours

Nom Prénom *Titre*	1ère insc. *Directeur*	Situation *Co-encadrant*
KAMENNOFF Nicolas *Utilisation d'une simulation 3D pour la compréhension de textes*	10/01/07 *D. Kayser*	bourse Epitech
BOSSARD Aurélien *Evolution et fusion d'informations extraites de données en langage naturel*	10/01/06 *D. Kayser*	Alloc moniteur (06-09) *T. Poibeau*
BOUTOUHAMI Sarah	11/01/06	vacation ANR

Argumentation — *D. Kayser*

MONDARY Thibault — 10/01/06 — Alloc moniteur (06-09)
Conception d'une méthode coopérative de mise à jour de règles d'extraction à base de corpus textuels — *A. Nazarenko*

MESSIANT Cédric — 10/01/06 — Alloc DGA moniteur (06-09)
Acquisition semi-automatique de patrons d'extraction — *A. Nazarenko* — *T. Poibeau*

Formation par la recherche

HAMON Olivier — 12/01/05 — CIFRE (Elda)
Méthodologies et protocoles d'évaluation des tyechnologies de la langue et de la communication homme-machine — *A. Nazarenko*

CAILLIAU Frédérick — 03/01/05 — CIFRE Sinequa
Conception et mise en oeuvre d'un système de gestion de resources linguistiques — *A. Nazarenko* — *Claude de Loupy*

BOUFFIER Amanda — 10/01/04 — Alloc (04-07)
Evolution et fusion des informations extraites de corpus en langage naturel — *D. Kayser* — *T. Poibeau*

WEISSENBACHER Davy — 10/01/03 — Alloc moniteur (03-06)
Normalisation syntaxico-sémantique de textes pour faciliter l'extraction d'informations : collaboration entre les techniques de TAL et d'apprentissage — *C. Fouqueré* — *A. Nazarenko*

Thèses soutenues (depuis 2004)

Nom Prénom	Situation ultérieure	Soutenance	Mention
Titre		*Directeur*	*Co-directeur*
Jury : Jury			
BOURGOIS Laure	CDD Sénégal	06/19/07	TH
Exploitation d'ontologies pour la combinaison automatique de composants logiciels		*F. Lévy*	*Alexandre Delteil*
Jury : Philippe Dague (rapp), Christophe Fouqueré (pdt), Laurent Hénocque, Farouk Toumani (rap)			
NOUIOUA Farid	Post-doc LRI	04/25/07	
Extraction et utilisation de normes pour le raisonnement causal dans un corpus textuel		*D. Kayser*	
Jury : Gérard Sabah (rapp), Pierre Marquis (rapp), Jean-Paul Haton, Pascal Nicolas, A. Nazarenko			
AIT EL MEKKI Touria	MCF Angers (LERIA)	12/21/04	TH
Construction semi-automatique d'index de fin de livre		*D. Kayser*	*A. Nazarenko*
Jury : Béatrice Daille (rapp), Jean-Pierre Desclés (rapp), Jean Véronis			

Habilitations à Diriger des Recherches soutenues (depuis 2004)

Nom Prénom — Soutenance

Titre

Jury :

Biébow Brigitte — 07/09/04

De Dasert à Terminæ ; voyage au pays de la modélisation des connaissances à partir de textes

Jury : Jean Charlet (rapp), Rose Dieng-Kuntz (rapp), Gilles Kassel (rapp), Daniel Kayser, François Lévy

Perrin-Nazarenko Adeline — 12/10/04

Donner accès au contenu des documents textuels. Acquisition de connaissances et analyse de corpus spécialisés

Jury : Maria Teresa Pazienza (rapp), Jean-Marie Pierrel (rapp), Bernard Victorri (rapp), Benoît Habert, Daniel Kayser, François Lévy

Chapitre 2
Organisation de conférences

Le LIPN a soutenu ou organisé les conférences et manifestations suivantes :

- *Implicit Computational Complexity and Logic*, septembre 2004, université Paris 13 (ACI GEOCAL), 30 participants, dont 8 étrangers.
- *Workshop on Implicit Computational Complexity*, 13-17/2/2006, (GEOCAL, *Geometry of Computation*, 2006), Marseille. 20 exposés et 50 participants (de 7 pays).
- *2es Journées thématiques "Apprentissage Artificiel & Fouille de Données"*, 27-28/4/2006, http://www-lipn.univ-paris13.fr/A3/AAFD06/
- *Journées ALEA*, 2005, (CIRM, SMF-CNRS, Luminy)
- *Marches aléatoires discrètes : théorie et applications en combinatoire, biologie algorithmique, informatique théorique, probabilités, physique statistique* (DRW-2003), Institut Henri Poincaré, 2003

Partie III
Publications

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

Articles dans des revues avec comité de lecture (ACL)

Articles dans des revues internationales avec comité de lecture sélectif

2007 [RI-6] Sylvie Després et Sylvie Szulman. Merging of Legal Micro-Ontologies from European Directives. *Artificial Intelligence and Law*, 15(2) :187–200, Feb. 2007.

[RI-5*] Thierry Poibeau. Semantic annotation : Mapping text to ontologies. *International Journal of Metadata Semantics and Ontologies*, 2(2), 2007.

2005 [RI-4] Daniel Kayser et Farid Nouioua. About norms and causes. *International Journal on Artificial Intelligence Tools*, 14 :7–23, 2005.

[RI-3] Adeline Nazarenko et Touria Aït El Mekki. Building back-of-the-book indexes. *Terminology*, 11(1) :193–218, 2005.

2004 [RI-2] Marie-Odile Cordier, Philippe Dague, François Lévy, Jacky Montmain, Marcel Staroswiecki et Louise Travé-Massuyès. Conflicts versus analytical redundancy relations : a comparative analysis of the model based diagnosis approach from the artificial intelligence and automatic control perspective. *IEEE Transactions on Systems, Man and Cybernetics*, 34 :2163–2177, 2004.

2003 [RI-1] Daniel Kayser. Abstraction and natural language semantics. *Philosophical Transactions of the Royal Society of London*, 358 :1261–1268, 2003.

Articles dans des revues nationales ou internationales

2008 [RE-16*] Sylvie Després et Hélène Richy. Métadonnées, ontologies et documents numériques. *Revue des Techniques de l'Ingénieur*, 2008.

2007 [RE-15*] Sophie Aubin, Julien Derivière, Thierry Hamon, Adeline Nazarenko, Thierry Poibeau et Davy Weissenbacher. Une infrastructure pour l'annotation linguistique de documents issus du web : le projet ALVIS. *Revue des Nouvelles Technologies de l'Information*, 2007.

[RE-14] Davy Weissenbacher. Les réseaux bayésiens : un formalisme adapté au traitement automatique des langues ? *Intelligence Artificielle*, 21(3) :371–389, Juin 2007.

2006 [RE-13*] Davy Weissenbacher. Les réseaux bayésiens : un formalisme adapté au traitement automatique des langues ? *Revue d'Intelligence Artificielle, Numéro spécial : "Modèles Graphiques Probabilistes"*, 2006. À paraître.

[RE-12] Haifa Zargayouna, Touria Ait El Mekki, Laurent Audibert et Adeline Nazarenko. IndDoc : An aid for the back-of-the-book indexer. *The Indexer*, 25(2) :122–125, 2006.

2005 [RE-11] Sylvie Després et Sylvie Szulman. Construction d'une ontologie formelle à partir d'un texte de droit communautaire. *International journal of theoretical and applied issues in specialized communication*, 11(2) :261–281, 2005.

[RE-10] Catherine Recanati. Raisonner avec des diagrammes : perspectives cognitives et computationnelles. *Intellectica*, 40 :9–42, 2005.

2004 [RE-9] Touria Aït El Mekki et Adeline Nazarenko. L'index de fin de livre, une forme de résumé indicatif ? *Traitement Automatique des Langues*, 45(1) :121–150, 2004.

[RE-8] Natalia Grabar et Thierry Hamon. Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'Intelligence Artificielle*, 18/1 :57–85, 2004.

[RE-7] Gaston Gross et Adeline Nazarenko. Quand la langue cause : paramètres d'une étude linguistique de la causalité. *Intellectica*, 38(1) :15–41, 2004.

[RE-6] Daniel Kayser et François Lévy. Modélisations symboliques du raisonnement causal. *Intellectica*, 38 :291–323, 2004.

[RE-5] François Lévy. Statistiques et contrefactuels, quelques explications. *Intellectica*, 38 :187–191, 2004. Introduction à "Expérimentation, inférence statistique et analyse causale" de B. Lecoutre dans le même numéro.

[RE-4] Adeline Nazarenko. Primitives logiques et réseaux sémantiques : quel bilan pour la terminologie ? *Revue d'Intelligence Artificielle*, 18(1) :41–56, 2004.

[RE-3] Thierry Poibeau. Annotation d'informations textuelles : le cas du Web sémantique. *Revue d'Intelligence Artificielle*, 18(1) :139–157, 2004.

2003 [RE-2] Daniel Kayser. Des modèles de nature différente peuvent approcher une même réalité ; réponse à Giuseppe Longo. *Intellectica*, 35 :175–177, 2003.

[RE-1] Daniel Kayser. Quelle place pour l'Intelligence Artificielle ? *lettre du département SHS du CNRS*, 67 :11–12, 2003.

Conférences invitées (INV)

Communications invitées dans une conférence internationale

2004 [IN-2] Thierry Poibeau. Content annotation for the semantic web. Sergio Bolasco, éditeur, *The NEMIS 2004 Final Conference*, pages 1–15, Rome, Italie, 2004. Universita di Roma1 « La Sapienza ». Conférence invitée lors de la conférence de clôture Nemis. Reprise ensuite en volume chez Springer.

2003 [IN-1] Daniel Kayser. Une approche inférentielle de la sémantique du langage naturel. *6th International Symposium on Programming and Systems*, page 9, Alger, Algérie, 2003.

Communications avec actes (ACT)

Communications dans des conférences internationales avec comité de lecture

2007 [CI-36*] Amanda Bouffier. From Texts to Structured Documents : The Case of Health Practice Guidelines. *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*, South Korea, Nov. 2007.

[CI-35*] Amanda Bouffier et Thierry Poibeau. Automatically Restructuring Practice Guidelines using the GEM DTD. Kevin Bretonnel Cohen, Dina Demner Fushman, Carol Friedman, Lynette Hirschman et John Pestian, éditeurs, *Proceedings of Biomedical Natural Language Processing (BioNLP'07)*, Prague, République Tchèque, 2007. ACL.

[CI-34*] Amanda Bouffier et Thierry Poibeau. Restructuring documents : A discourse based approach. Bulgarie, Oct. 2007.

[CI-33] Laure Bourgois, Alexandre Delteil et François Lévy. A Formalism for Web Services Properties. *Proceedinfgs of the International Conference on Internet and Web Applications and Services (ICIW'07)*, page 28, 05 2007.

[CI-32*] Sylvie Després et Valentina Ceausu. 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 2007 + ASWC 2007)*, Corée, 11 2007.

[CI-31] Sylvie Després et Sonia Mhiri. Ontology Visualization Tool for Indexing DICOM Structured Reporting (SR) Documents. *Proceedings of the 10th International Protégé Conference*, pages 150–153, Hongrie, 07 2007.

[CI-30*] Sylvie Després et Sonia Mhiri. Usability via a Visualization Tool for the Semantic Indexing of Medical Reports (DICOM SR). *Proceedings of USAB 07 (Usability & HCI for Medicine and Health Care3)*, Autriche, 11 2007.

[CI-29*] Thierry Hamon, Julien Derivière et Adeline Nazarenko. Ogmios : a scalable NLP platform for annotating large web document collections. Birmingham, UK, Jul. 2007. Electronic publication. ALVIS STREP Project.

[CI-28] Thierry Hamon, Adeline Nazarenko, Thierry Poibeau, Sophie Aubin et Julien Derivière. A Robust Linguistic Platform for Efficient and Domain specific Web Content Analysis. Pittsburgh, USA, May 2007. ALVIS STREP Project.

[CI-27] Thibault Mondary, Amanda Bouffier et Adeline Nazarenko. Between browsing and search, a new model for navigating through large documents. pages 634–639, Delphi, Grèce, May 2007. Lawrence Erlbaum Associates.

[CI-26] Catherine Recanati. Characteristics of diagrammatic reasoning. pages 510–515, Delphi, Grèce, May 2007. Lawrence Erlbaum Associates.

[CI-25*] Catherine Recanati, Nicoleta Rogovschi et Younès Bennani. Sequencing of verbs - a study on tense and aspect using unsupervised learning. Borovets, Bulgary, Sep. 2007. Institute for Parallel Processing, Bulgarian Academy of Sciences.

[CI-24*] Catherine Recanati, Nicoleta Rogovschi et Younès Bennani. The structure of verbal sequences analyzed with unsupervised learning techniques. Poznań, Poland, Oct. 2007.

[CI-23*] Davy Weissenbacher et Adeline Nazarenko. A Bayesian approach combining surface clues and linguistic knowledge : Application to the anaphora resolution problem. Borovets, Bulgary, Sep. 2007. Institute for Parallel Processing, Bulgarian Academy of Sciences.

2006 [CI-22] Sophie Aubin et Thierry Hamon. Improving Term Extraction with terminological resources. Tapio Salakoski, Filip Ginter, Sampo Pyysalo et Tapio Pahikkala, éditeurs, *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, LNAI 4139, pages 380–387, 2006.

[CI-21] Laure Bourgois. Finding alternative Web services to Parry Breakdowns. *Proceedings of the Third Starting AI Researchers' Symposium*, pages 232–243. IOS Press, 2006.

[CI-20] Laure Bourgois, Alexandre Delteil et François Lévy. Web Services subsumption with a specific PDL. *Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT/ ICIW)*, page 158. IEEE Computer Society, 2006.

[CI-19] Julien Derivière, Thierry Hamon et Adeline Nazarenko. A scalable and distributed NLP architecture for Web document annotation. Tapio Salakoski, Filip Ginter, Sampo Pyysalo et Tapio Pahikkala, éditeurs, *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, LNAI 4139, pages 56–67, 2006.

[CI-18] Julien Derivière, Thierry Hamon, Adeline Nazarenko, Thierry Poibeau et Davy Weissenbacher. A robust linguistic infrastructure for efficient web content analysis : the ALVIS project. *Symposium on Digital Semantic Content across Cultures*, Paris, France, 2006. 15 pages.

[CI-17] Sylvie Després et Sylvie Szulman. Terminae Method and Integration Process for Legal Ontology Building. *Advances in Applied Artificial Intelligence : 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2006, Annecy, France, June 27-30, 2006.*, pages 1014–1023, Annecy, France, June 2006. Springer Berlin/Heidelberg.

[CI-16] Natalia Grabar et Thierry Hamon. Terminology structuring through the derivational morphology. Tapio Salakoski, Filip Ginter, Sampo Pyysalo et Tapio Pahikkala, éditeurs, *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, LNAI 4139, pages 652–663, 2006.

[CI-15] Adeline Nazarenko, Erick Alphonse, Julien Derivière, Thierry Hamon, Guillaume Vauvert et Davy Weissenbacher. The ALVIS format for linguistically annotated documents. *Proceedings of the 5th international conference on Language Resources and Evaluation, LREC 2006*, pages 1782–1786. ELDA, 2006.

[CI-14] Thierry Poibeau. Dealing with metonymic readings of Named Entities. *The 28th Annual Conference of the Cognitive Science Society (CogSci 2006)*, pages 1962–1968, Vancouver, Canada, 2006. Cognitive Science Society.

[CI-13] Sampo Pyysalo, Tapio Salakoski, Sophie Aubin et Adeline Nazarenko. Lexical adaptation of Link Grammar to the biomedical sublanguage : a comparative evaluation of three approaches. *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006)*, pages 60–67. Sophia Ananiadou, Juliane Fluck, 2006.

[CI-12] Davy Weissenbacher. Bayesian Network, a model for NLP ? *European Conference of the Association for Computational Linguistics (EACL 2006)*, pages 195–198, 2006.

2005 [CI-11] Touria Aït El Mekki et Adeline Nazarenko. Using NLP to build the hypertextuel network of a back-of-the-book index. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 316–320. Institute for Parallel Processing, Bulgarian Academy of Sciences, 2005.

[CI-10] Sophie Aubin, Adeline Nazarenko et Claire Nédellec. Adapting a general parser to a sublanguage. 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'05)*, pages 89–93. INCOMA Ltd., 2005.

[CI-9] Sylvie Després et Sylvie Szulman. Merging of Legal Micro-Ontologies from European Directives. *LOAIT - Legal Ontologies and Artificial Intelligences Techniques*, pages 149–155. IAAIL Workshop series, 2005.

[CI-8] Daniel Kayser. What logic can and cannot do in natural language semantics. *First Indian Conference on Logic and its relationship with other disciplines*, Actes en cours d'édition, Bombay, Inde, 2005.

2004 [CI-7] Sylvie Després et Sylvie Szulman. Construction of a Legal Ontology from a European Community Legislative Text. Thomas Gordon, éditeur, *Jurix 2004- Berlin*, pages 79–88. IOS Press, 2004.

[CI-6] Françoise Gayral, Daniel Kayser et François Lévy. Challenging the Principle of Compositionality in Interpreting Natural Language Texts. *Conference on Compositionality, Concepts and Cognition*, pages 111–127, Düsseldorf, Allemagne, 2004.

[CI-5] Daniel Kayser et Farid Nouioua. About norms and causes. Valerie Barr et Zadravko Markov, éditeurs, *The 17th Florida Artificial Intelligence Research Society Conference (FLAIRS'04)*, pages 502–507. AAAI Press, 2004.

[CI-4] Daniel Kayser et Farid Nouioua. Representing knowledge about norms. *Proc. 16th European Conference on Artificial Intelligence*, Valencia, Spain, pages 363–367. IOS Press, 2004.

[CI-3] Thierry Poibeau. Automatic extraction of paraphrastic phrases from medium size corpora. Sergei Nirenburg, éditeur, *Actes de la conférence Computational Linguistics (COLING'2004)*, pages 638–644. Association for Computational Linguistics, 2004.

[CI-2] Thierry Poibeau et Bénédicte Goujon. Semi-automatic generation of command grammar. *Language and Resource Evaluation Conference (LREC 2004)*, pages 2147–2150, Lisbonne, Portugal, 2004.

2003 [CI-1] Thierry Poibeau. The multilingual Named Entity recognition framework. Ann Copestake et Jan Hajic, éditeurs, *European Conference of the Association for Computational Linguistics (EACL 2003)*, pages 155–158, Budapest, Hongrie, 2003. Association for Computational Linguistics.

Workshops et autres conférences avec comité de lecture

2008 [CO-64*] Sylvie Després. Construction dï¿½ontologies à partir de textes dans le domaine juridique. *Actes du colloque annuel de la Société Française de Terminologie (SFT)*, France, 2008.

2007 [CO-63] Laurent Audibert. Désambiguïsation lexicale automatique : sélection automatique d'indices. pages 13–22, Toulouse, France, Juin 2007.

[CO-62] Aurélien Bossard. Vers une ressource prédicative pour l'extraction d'information. page 154, Toulouse, France, Juin 2007.

[CO-61*] Amanda Bouffier et Thierry Poibeau. Exploiter la structure des documents : un enjeu pour la fouille de textes. *Actes du colloque Veille Stratégique Scientifique & Technologique (VSST'07)*, Maroc, Oct. 2007.

[CO-60*] Valentina Ceausu et Sylvie 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)*, Canada, 10 2007.

[CO-59] Sylvie Despres, Frédéric Fürst et Sylvie Szulman. Construction d'une ontologie du domaine HSE. *Ingénierie des Connaissances - 18èmes journées francophones*, pages 133–144, Grenoble, France, 2007. Cépaduès.

[CO-58] Sylvie Després, Sonia Mhiri et Ezzaddine Zagrouba. Unification d'ontologies hétérogènes pour l'indexation sémantique de comptes rendus standardisés d'examens du patient. *Actes de l'atelier Ontologies et gestion de l'hétérogénéité sémantique (OGHS). Plateforme AFIA 2007*, pages 60–62, Grenoble, France, 07 2007.

[CO-57] Thierry Hamon, Julien Derivière et Adeline Nazarenko. Ogmios : une plate-forme d'annotation linguistique. pages 103–112, Toulouse, France, Juin 2007. ALVIS STREP Project.

[CO-56] Luc Lamontagne, Rim Bentibebel, Erwan Miry et Sylvie 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, 08 2007.

[CO-55] Jean-Vincent Loddo et Thierry Hamon. Marionnet/Iutoppix : un environnement pédagogique pour la simulation de réseaux locaux. *Environnement Informatique pour l'Apprentissage Humain (EIAH'07)*, Juil. 2007. Démonstration.

[CO-54*] Thierry Poibeau. Up13 : Knowledge poor methods (sometimes) perform poorly. *Proceedings of the ACL SemEval workshop ACL SemEval (Semantic Evaluation) workshop*, Tchèque, République Prague, 2007. Association for Computational Linguistics.

[CO-53*] Thierry Poibeau et Aurélien Bossard. Adaptation d'une ressource prédicative pour l'extraction d'information. *Actes du Colloque Lexique et Grammaire 2007 Colloque Lexique et Grammaire 2007*, Bonifacio, France, 2007. LADL.

[CO-52] Catherine Recanati et Nicoleta Rogovschi. Enchaînements verbaux - étude sur le temps et l'aspect utilisant des techniques d'apprentissage non supervisé. pages 379–388, Toulouse, France, Juin 2007.

[CO-51] Nicoleta Rogovschi, Younès Bennani et Catherine Recanati. Apprentissage neuro-markovien pour la classification non supervisée de données structurées en séquences. *Actes de Extraction et Gestion des Connaissances (EGC'07)*, Atelier « Fouille de données et algorithmes biomimétiques », pages 49–60, Namur, Belgique, Jan. 2007. Cépaduès Editions.

[CO-50*] Sylvie Szulman et Sylvie Despres. Exploitation linguistique d'une ontologie générique. *TIA 2007 Terminologie et Intelligence artificielle*, Sophia-Antipolis, France, 10 2007.

[CO-49] Davy Weissenbacher et Adeline Nazarenko. A bayesian classifier for the recognition of the impersonal occurrences of the 'it' pronoun. pages pp.145–150, Lagos, Portugal, May 2007.

[CO-48] Davy Weissenbacher et Adeline Nazarenko. Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. pages 145–155, Toulouse, France, Juin 2007.

[CO-47*] Haifa Zargayouna, Olivier Hamon et Adeline Nazarenko. Evaluation des outils terminologiques : état des lieux et propositions. *Actes de la 7ème Conférence : Terminologie et Intelligence Artificielle (TIA'07)*, Sophia Antipolis, France, 10 2007.

[CO-46] Haifa Zargayouna, Victor Rosas et Sylvie Salotti. Shallow parsing of INEX Queries. *Comparative Evaluation of XML Information Retrieval Systems 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, pages 284–293, Dagstuhl, Deutschland, 02 2007. LNCS, Volume Number 4518.

2006

[CO-45] Touria Aït El Mekki et Adeline Nazarenko. An application-oriented terminology evaluation : the case of back-of-the book indexes. Rute Costa, Fidélia Ibekwe-SanJuan, Susanne Lervad, Marie-Claude L'Homme, Adeline Nazarenko et Henrik Nilsson, éditeurs, *Workshop on Terminology design : quality criteria and evaluation methods (TermEval)*, pages 18–21, Genoa, Italie, 2006.

[CO-44] Sophie Aubin, Julien Derivière, Thierry Hamon, Adeline Nazarenko, Thierry Poibeau et Davy Weissenbacher. ALVIS : A robust linguistic infrastructure for efficient semantic content analysis. *Symposium on Digital Semantic Content across Cultures*, Paris, France, 2006.

[CO-43] Amanda Bouffier et Thierry Poibeau. Segmentation et structuration de textes procéduraux pour l'aide à la modélisation de connaissances : le rôle de la structure visuelle. *Semaine de la Connaissance*, pages 195–205, Nantes, France, 2006.

[CO-42] Laure Bourgois, Alexandre Delteil et François Levy. Réparation de Web services composites. *Atelier Conception des systèmes d'information et services Web, INFORSID'06*, 2006. 12 pages, //sisw06.loria.fr/SISW$_$files/articles/BourgoisArticle.pdf.

[CO-41] Frederik Cailliau. Unifier la gestion de ressources linguistiques en contexte multilingue. *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2006*, pages 455–461. Presses universitaires de Louvain, 2006.

[CO-40] Julien Derivière et Thierry Hamon. Annotation linguistique de documents Web dans une architecture distribuée et adaptable. *The French Perl Workshop (FPW2006)*, Nov. 2006. Communication orale, ALVIS STREP project.

[CO-39] Daniel Kayser et Farid Nouioua. Une expérience de sémantique inférentielle. *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2006*, pages 246–255. Presses Universitaires de Louvain, 2006.

[CO-38] Adeline Nazarenko, Claire Nédellec, Erik Alphonse, Sophie Aubin, Thierry Hamon et Alain-Pierre Manine. Semantic annotation in the ALVIS project. *Proceeding of IIIA-2006 : International Workshop on Intelligent Information Access*, Helsinki, Finlande, 2006.

[CO-37] Farid Nouioua et Pascal Nicolas. Using answer set programming in an inference-based approach to natural language semantics. Alexander Koller Johan Bos, éditeur, *Inference in Computational Semantics ICoS-5*, pages 77–86, Buxton, France, 2006.

[CO-36] Thierry Poibeau. Linguistically grounded models of language change. Ron Sun, éditeur, *The 28th Annual Conference of the Cognitive Science Society (CogSci 2006)*, page 2587, Vancouver, Canada, 2006.

[CO-35] Haifa Zargayouna et Nejla Amara. LiteMap : An ontology mapping approach for Mobile Agents' context-awareness. *OTM Workshops 2006 International Workshop on Context-Aware Mobile Systems*, pages 1934–1943, Montpellier, France, 2006. LNCS 4278.

2005 [CO-34] Touria Aït El Mekki et Adeline Nazarenko. Histoire et avenir des index de fin de livre. *Colloque international Indice, Index, Indexation*, Lille, France, 2005.

[CO-33] Touria Aït El Mekki et Adeline Nazarenko. ï¿½valuer le réseau terminologique d'un index de document. *Rencontres Terminologie et Intelligence Artificielle ; Atelier Evaluation des outils terminologiques*, Rouen, France, 2005.

[CO-32] Thierry Delbecque, Pierre Zweigenbaum, Jean-François Berroyer et Thierry Poibeau. Le système STIM/LIPN à EQueR 2004, tâche médicale. *Atelier EQueR (conférence TALN'2005)*, pages 89–94, Dourdan, France, 2005.

[CO-31] Julien Derivière, Thierry Hamon, Adeline Nazarenko et Guillaume Vauvert. Développement d'une plate-forme d'enrichissement des documents textuels : l'expérience du projet ALVIS. *Journée ATALA "Articuler les traitements sur corpus"*, Paris, France, 2005.

[CO-30] Sylvie Després et Sylvie Szulman. Construction d'une ontologie du droit documentaire. *IC'2005*, pages 85–96. Presses Universitaires de Grenoble, 2005.

[CO-29] Thierry Hamon. Indexer les documents spécialisés : les ressources terminologiques contrôlées sont-elles suffisantes ? *Actes de Terminologie et Intelligence Artificielle 2005*, pages 71–82. PUR, 2005.

[CO-28] Goritsa Ninova, Adeline Nazarenko, Thierry Hamon et Sylvie Szulman. Comment mesurer la couverture d'une ressource terminologique pour un corpus ? *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2005*, pages 293–302. ATALA, 2005.

[CO-27] Farid Nouioua. Norm based causal reasoning in textual corpus. *Proceedings of the Sixth International Workshop on Computational Semantics IWCS-6*, pages 396–400, Tilburg, France, 2005.

[CO-26] Farid Nouioua. Raisonnement stratifié à base de normes pour inférer les causes dans un corpus textuel. *The Seventh International Symposium On Programming and Systems*, pages 81–92. USTHB d'Alger, 2005.

[CO-25] Farid Nouioua. Why did the accident happen ? A norm-based reasoning approach. *Logical Aspects of Computational Linguistics, student session*, pages 31–34. Université de Bordeaux, 2005.

[CO-24] Thierry Poibeau. Architectures articulant des représentations hétérogènes : l'exemple de Gate. B. Habert, A. Salem et S. Heiden, éditeurs, *Journée ATALA (Association pour le Traitement Automatique des Langues)*, Paris, France, 2005. Atala. www.limsi.fr/\-Individu/\-habert/\-Projets/\-JourneeATALA05ArticulerLesTraitements/\-Poibeau.ppt.

[CO-23] Thierry Poibeau. Parcours interprétatifs et terminologie. Yannick Toussaint, éditeur, *Actes du colloque Terminologie et Intelligence Artificielle*, pages 12–25. Université de Rouen, 2005.

[CO-22] Thierry Poibeau. Sur le statut référentiel des entités nommées. Michèle Jardino, éditeur, *Conférence Traitement Automatique des Langues 2005*, pages 173–183, Dourdan, France, 2005.

[CO-21] Davy Weissenbacher. A Bayesian Network for the resolution of non-anaphoric pronoun It. *Bayesian Methods for Natural Language Processing Workshop at NIPS*, Canada, 2005.

2004

[CO-20] Erick Alphonse, Sophie Aubin, Philippe Bessières, Gilles Bisson, Thierry Hamon, Sandrine Lagarrigue, Alain-Pierre Manine, Adeline Nazarenko, Claire Nédellec, Mohamed Ould Abdel Vetah, Thierry Poibeau et Davy Weissenbacher. Extraction d'information appliquée au domaine biomédical : apprentissage et traitement automatique de la langue. *Actes de la Conférence Internationale sur la Fouille de textes (CIFT'04)*, pages 7–19, La Rochelle, France, 2004.

[CO-19] Erick Alphonse, Sophie Aubin, Philippe Bessières, Gilles Bisson, Thierry Hamon, Sandrine Lagarrigue, Adeline Nazarenko, Alain-Pierre Manine, Claire Nédellec, Mohamed Ould Abdel Vetah, Thierry Poibeau et Davy Weissenbacher. Event-based information extraction for the biomedical domain : the Caderige project. Nigel Collier, Patrick Ruch et Adeline Nazarenko, éditeurs, *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (COLING'04)*, pages 43–49, Geneva, Suisse, 2004.

[CO-18] Jean Charlet, Touria Aït El Mekki, Didier Bourigault, Adeline Nazarenko, Régine Teulier et Baruk Toledano. CEDERILIC : constitution d'un livre et d'un index numériques. *Actes de 7ème conférence Internationale sur le Document numérique (CIDE'04)*, pages 187–204. Semaine du document numérique, 2004.

[CO-17] Françoise Gayral, Daniel Kayser et François Lévy. The role of time in considering collections. *Journées de Sémantique et Modélisation*, pages 30–34, Lyon, France, 2004.

[CO-16] Natalia Grabar et Thierry Hamon. Repérage de relations sémantiques entre termes : sur la piste de la morphologie. *Actes de Journées francophones d'Ingénierie des Connaissances (IC'2004)*, pages 237–248. Presses Universitaires de Grenoble, 2004.

[CO-15] Thierry Hamon. Plateforme d'annotation terminologique de corpus des résumés d'articles de biologie. *Actes des 15ièmes journées francophones d'Ingénierie des Connaissances (IC'2004)*, Lyon, France, 2004. Presses universitaires de Grenoble. Poster.

[CO-14] Adeline Nazarenko et Touria Aït El Mekki. Une mesure de pertinence pour le tri de l'information dans un index de fin de livre. *Actes de la 11ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2004*, pages 1–10. ATALA, 2004.

[CO-13] Thierry Poibeau. Pré-analyse de corpus. G. Purnelle, C. Fairon et A. Dister, éditeurs, *Journées d'Analyse des Données Textuelles (JADT 2004)*, pages 897–904, Louvain, Belgique, 2004. Presses Universitaires de Louvain.

[CO-12] Thierry Poibeau et Bénédicte Goujon. Acquisition semi-automatique de grammaires de commande. I. Marlien et B. Bel, éditeurs, *Journées d'etude de la Parole (JEP'2004)*, pages 405–409, Fès, Maroc, 2004. Association Francophone de la Communication Parlée. Collaboration avec Thales Recherche et Technologie.

[CO-11] Sylvie Szulman et Brigitte Biébow. OWL et Terminae. *IC'2004 - 15ème journées francophones d'Ingénierie des Connaissances*, pages 41–52. Presses Universitaires de Grenoble, 2004.

[CO-10] Davy Weissenbacher. La relation de synonymie en génomique. *Actes des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, pages 298–303. ATALA, 2004.

[CO-9] Haifa Zargayouna et Sylvie Salotti. Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. *15ème journées francophones d'Ingénierie des Connaissances (IC'2004)*, pages 249–260. Presses Universitaires de Grenoble, 2004.

[CO-8] Haifa Zargayouna et Sylvie Salotti. Mesure de similarité sémantique pour l'indexation de documents semi-structurés. *12ème Atelier de Raisonnement à Partir de Cas*, pages 19–29, Villetaneuse, France, 2004. LIPN.

[CO-7] Haifa Zargayouna et Sylvie Salotti. Semindex : a model of semantic indexing on XML documents. *26th European Conference on Information Retrieval (ECIR'2004)*, 2004. Poster.

2003

[CO-6] Touria Aït El Mekki et Adeline Nazarenko. Le réseau terminologique, un élément central pour la construction d'index de document. *Actes des 5èmes rencontres Terminologie et Intelligence Artificielle (TIA'03)*, pages 1–10, Strasbourg, France, 2003.

[CO-5] Touria Aït El Mekki et Adeline Nazarenko. Semi-automatic method for the construction of a document index. *Workshop :"Applied Natural Language Processing, possible applications for the Semantic Web" associated with the EUROLAN-2003 summer school*, Bucarest, Roumanie, 2003.

[CO-4] Nathalie Aussenac-Gilles, Brigitte Biébow et Sylvie Szulman. D'une méthode à un guide pratique de modélisation de connaissances à partir de textes. *Actes des 5èmes rencontres Terminologie et Intelligence Artificielle (TIA'03)*, pages 41–53. LIIA - ENSAIS, 2003.

[CO-3] Françoise Gayral, Daniel Kayser et François Lévy. Collections évolutives et partitions. *Colloque "La partition en langue et en discours"*, France, 2003.

[CO-2] Adeline Nazarenko et Touria Aït El Mekki. Une mesure de pertinence pour le tri de l'information dans un document. *Atelier Acquisition, apprentissage et exploitation de connaissances sémantiques pour l'accès au contenu textuel, Plateforme AFIA*, Laval, France, 2003.

[CO-1] Haifa Zargayouna et Sylvie Salotti. Indexation sémantique de documents XML. *Extraction et gestion des connaissances (EGC'2003). Atelier : Fouille de données et recherche d'informations dans des bases de données multimédia semi-structurées*, 2003. Poster.

Direction d'ouvrages (DO)

Livres

2004 [LI-2] Daniel Kayser, Adina Magda Florea et Stefan Gheorghe Pentiuc. *Intelligence artificielle et agents intelligents*. Printech, 2004, ISBN 973-652-976.

2003 [LI-1] Thierry Poibeau. *Extraction automatique d'information : Du texte brut au Web sémantique*. Lavoisier, 2003, ISBN 2-7462-0610-2.

Chapitres de livre

2008 [CL-11*] Sylvie Szulman, Nathalie Aussenac Gilles et Sylvie Despres. 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*. IOS press, 2008.

2007 [CL-10] Adeline Nazarenko et Touria Aït El Mekki. Building back-of-the-book indexes. Fidelia Ibekwe SanJuan, Anne Condamines et Maria Teresa Cabré, éditeurs, *Application-Driven Terminology Engineering*, Current Topics 2, pages 179–201. John Benjamins, 2007.

[CL-9] Antoine Rozenknop et Jean-Cédric Chappelier. Syntactic Processing. Martin Rajman et Vincenzo Pallota, éditeurs, *Speech and Language Engineering*, pages 177–204. Presses Polytechniques et Universitaires Romandes (CH)/ Marcel Dekker (US), 2007.

2006 [CL-8] Adeline Nazarenko. *Le point sur l'état actuel des connaissances en Traitement Automatique des Langues*, Gérard Sabah, éditeur, *Compréhension des langues et interaction*, pages 31–70. 2006.

2005 [CL-7] Brigitte Biébow, Sylvie Szulman et Nathalie Aussenac-Gilles. *Modélisation du domaine par une méthode fondée sur l'analyse de corpus*, R. Teulier ; J. Charlet ; P. Tchounikine, éditeur, *Ingénierie des connaissances*, pages 49–71. L'Harmattan, 2005.

[CL-6] Françoise Gayral, Daniel Kayser et François Lévy. *Challenging the Principle of Compositionality in Interpreting Natural Language Texts*, E. Machery, M. Werning et G. Schurz, éditeurs, *The Compositionality of Meaning and Content*, volume 2, pages 83–105. Ontos Verlag, 2005.

[CL-5] Adeline Nazarenko. *Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ?*, Anne Condamines, éditeur, *Sémantique et corpus*, pages 211–244. Traité IC2, série Cognition et traitement de l'information. Lavoisier, 2005.

[CL-4] Claire Nédellec et Adeline Nazarenko. *Ontology and Information Extraction : a necessary symbiosis*, Bernardo Magnini Paul Buitelaar, Philipp Cimiano, éditeur, *Ontology Design and Population*, pages 155–170. IOS Press, 2005.

[CL-3] Thierry Poibeau. *Content annotation for the semantic Web*, Spiros Sirmakessis, éditeur, *Knowledge Mining*, pages 250–262. 2005. ISBN 3-540-25070-0.

[CL-2] Thierry Poibeau. *Una metodologica per l'annotazione e l'estrazione d'informazione*, S. Bolasco, A. Canzonetti et F.M. Capo, éditeurs, *Knowledge Mining : Uno strumento strategico per imprese e institutioni*, pages 37–45. ASU, 2005. ISBN 88-7975-341-X.

2004 [CL-1] Adeline Nazarenko et Thierry Poibeau. *L'évaluation des systèmes d'analyse et de compréhension de textes*, Stéphane Chaudiron, éditeur, *L'évaluation des systèmes de traitement de l'information*, pages 127–148. 2004. ISBN 2-7462-0862-8.

Rapports de contrats

2006 [RC-9] Erick Alphonse, Sophie Aubin, Julien Derivière, Thierry Hamon, Dunja Mladenic, Adeline Nazarenko, Claire Nédellec, Thierry Poibeau, Davy Weissenbacher et Qiang Zhou. Deliverable D5.1 : Report on method and language for the production of the augmented document representations. ALVIS, 2006.

[RC-8] Gérard Bécher, Patrice Enjalbert, Estelle Fievé, Laurent Gosselin, François Lévy et Gérard Ligozat. Rapport technique du projet OGRE. 92 pages PI TCAN du CNRS, 2006.

[RC-7] Julien Derivière et Adeline Nazarenko. Deliverable D5.2 : Report on theory and software of normalization options for IR (platform conception). Technical report, ALVIS IST-1-002068 STREP project, 2006.

[RC-6] Thierry Poibeau et Adeline Nazarenko. Rapport InfoM@gic "État de l'art des méthodes linguistiques et statistiques pour l'extraction d'informations". 25 pages, Projet Infom@gic, 2006.

2004 [RC-5] Erick Alphonse, Philippe Bessières, Gaël Deral, Gaëtan Lehman, Alain-Pierre Manine, Claire Nédellec, Sophie Aubin, Julien Derivière, Thierry Hamon, Adeline Nazarenko, Thierry Poibeau et Davy Weissenbacher. Deliverable D6.2 : Requirements for integration of WP6 results into WP5 normalization and representation tasks and into WP9. ALVIS deliverable report ALVIS, 2004.

[RC-4] Sophie Aubin, Philippe Bessières, Thierry Hamon, Gaëtan Lehman, Alain-Pierre Manine, Adeline Nazarenko, Claire Nédellec, Mohamed Ould Abdel Vetah et Thierry Poibeau. Deliverable D6.1 : Requirement and training data for biological domain. ALVIS Deliverable Report ALVIS, 2004.

[RC-3] Sophie Aubin, Thierry Hamon et Adeline Nazarenko. Analyse terminologique (application et évaluation sur des textes de biologie). Contrat ExtraPloDocs (laboratoire LIPN, laboratoire INRA/MIG, Isoft), 2004.

[RC-2] Sophie Aubin et Adeline Nazarenko. Choix, évaluation et adaptation d'un analyseur syntaxique pour l'extraction d'information. Rapport ExtraPloDocs 3.3a. Contrat ExtraPloDocs (laboratoire LIPN, laboratoire INRA/MIG, Isoft), 2004.

[RC-1] Thierry Poibeau et Jean-François Berroyer. Analyse et repérage des entités nommées. 60 pages, rapport du Projet RNTL ExtraPloDocs 3.3b, 2004.

Rapports

2006 [Ra-3] Catherine Recanati, Younès Bennani et Nicoleta Rogovschi. Analyse de séquences verbales - une segmentation numérique des temps et catégories aspectuelles dans des récits d'accidents. 19 pages, 2006.

2005 [Ra-2] Claire Nédellec et Adeline Nazarenko. Ontologies and Information Extraction. LIPN Internal Report, 2005.

2004 [Ra-1] Catherine Recanati. Diagrammes pour résoudre le problème d'Einstein et celui d'un joueur de Mastermind. 29 pages, 2004.

Outils

2006 [Lo-7] Sophie Aubin et Thierry Hamon. YaTeA - version 2006. YaTeA (Yet Another Term ExtrActor) aims at identifying and extracting noun phrases which are potential terms (*i.e.* term candidates). Projet Européen STREP IST-1-002068-STP - ALVIS, 2006.

[Lo-6] Julien Derivière et Thierry Hamon. Alvis NLP platform. A scalable architecture using existing NLP tools to annotate large collections of web documents, 2006.

[Lo-5] Thierry Hamon. SynoTerm - version 2006. Logiciel de détection de relations de synonymie entre termes, 2006.

[Lo-4] Sylvie Szulman. Logiciel TERMINAE - version 2006. Plate-forme d'aide à la construction de terminologies et d'ontologies à partir de l'étude de textes. Intégre un concordancier (Linguae) et permet la visualisation des résultats de Lexter et/ou Syntex développés par D. Bourigault, 2006.

2005 [Lo-3] Laurent Audibert. Logiciel Concordanta. Logiciel de type concordancier, 2005.

[Lo-2] Laurent Audibert. Logiciel xexplora et explora. Outil de recherche et d'observation de phénomènes linguistiques dans les corpus écrits, 2005.

2004 [Lo-1] Thierry Poibeau et Jean-François Berroyer. Tagen, un système de reconnaissance d'entités nommées. 60 pages, 2004.

Miscellanea

2005 [Mi-1] Daniel Kayser. Compte-rendu de lecture du livre "Interpréter en contexte" sous la dir. de Francis Corblin et de Claire Gardent. *T.A.L.*, 46(1) :197–200, 2005.

Partie IV
Acronymes

A³ Apprentissage Artificiel et Applications

ACI Action Concertée Incitative

AM allocataire moniteur

ANR Agence National de la Recherche

ATER attaché temporaire d'enseignement et de recherche

CIFRE Conventions Industrielles de Formation par la Recherche

CNRS Centre National de la Recherche Scientifique

CR chargé de recherche

EA Equipe d'Accueil

FRE Formation de Recherche en Evolution

HDR Habilitation à Diriger des Recherches

IG Institut Galilée

IUTB IUT de Bobigny

IUTV IUT de Villetaneuse

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

LCR Logique, Calcul et Raisonnement

Lim&Bio Laboratoire d'Informatique Médicale et Bio-Informatique, université Paris 13 - EA

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

LLI Laboratoire de Linguistique-Informatique, université Paris 13 - FRE 2882

MCF maître de conférences

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

PPF Plan Pluri-Formation

PU professeur des universités

RCLN Représentation des Connaissances et Langage Naturel

RNTL Réseau National de recherche et d'innovation en Technologies Logicielles

SMF Société Mathématique de France

UE Union européenne

UMR Unité Mixte de Recherche