INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet AÏDA

Modélisation et Apprentissage pour l'Interprétation de Données et l'Aide à la décision

Rennes

THÈME 3A

Table des matières

1 Composition de l'équipe **3**

2 Présentation et objectifs généraux **4**
2.1 Présentation générale et objectifs 4

3 Fondements scientifiques **5**
3.1 Aide à la surveillance de systèmes physiques 5
3.2 Apprentissage automatique 8
3.2.1 Inférence grammaticale et programmation logique inductive 8
3.2.2 Classification 10
3.3 Recherche d'information dans un ensemble de documents, construction de lexiques 11
3.3.1 Recherche d'information - Indexation automatique 11
3.3.2 Analyse des séquences complexes 12
3.3.3 Acquisition automatique d'informations lexicales à partir de corpus . . . 13

4 Domaines d'applications **14**
4.1 Panorama 14
4.2 La génomique 14
4.3 Surveillance de systèmes physiques 15
4.4 La recherche d'information et l'accès à des bases de documents ou de services . 16
4.4.1 Recherche d'information 17
4.4.2 Système coopératif d'accès à un ensemble de services 17

5 Logiciels **18**
5.1 Unam : logiciel d'apprentissage inductif avec contraintes 18

6 Résultats nouveaux **18**
6.1 Modélisation de systèmes (ou d'activités complexes) évoluant dans le temps en vue de leur surveillance en ligne 18
6.1.1 Acquisition de scénarios à partir de modèles 19
6.1.2 Construction d'un automate diagnostiqueur 21
6.1.3 Surveillance de parcelles agricoles 22
6.2 Apprentissage automatique et structuration de données 23
6.2.1 Analyse de séquences 24
6.2.2 Inférence grammaticale 26
6.2.3 Apprentissage de structures d'arbres 27
6.2.4 Réduction de la complexité d'un système descriptif 28
6.3 Ingénierie de la langue 30
6.3.1 Grammaires minimalistes et logique linéaire 33
6.4 EIAO (Assistants intelligents pour l'enseignement) 34
6.4.1 Individualisation des logiciels de formation 34
6.4.2 Interaction dans les EIAO de calcul formel 34

6.5 Raisonnements et logiques non classiques 35
6.6 Planification et révision de croyances dans un système de dialogue 35

7 Contrats industriels (nationaux, européens et internationaux) 36
7.1 Modélisation, diagnostic et supervision de réseaux de télécommunication 36
7.2 Amélioration des traitements des informations temporelles dans les graphes causaux temporels 37
7.3 Inférence grammaticale régulière pour l'apprentissage de la syntaxe en reconnaissance de la parole 37
7.4 L'interaction dans les EIAO intégrant des instruments de calcul formel 37
7.5 Développement d'assistants intelligents au sein des logiciels de formation professionnelle 37
7.6 Définition et mise en œuvre d'une théorie de la révision des croyances dans le contexte d'un dialogue coopératif 38
7.7 Acquisition automatique d'éléments du Lexique Génératif pour améliorer les performances de systèmes de recherche d'information 38
7.8 Conception et contrôle de stimulateurs-défibrillateurs cardiaques intégrés 38

8 Actions régionales, nationales et internationales 39
8.1 Actions nationales 39
8.2 Réseaux et groupes de travail internationaux 39
8.3 Relations bilatérales internationales 39
8.4 Accueils de chercheurs étrangers 39

9 Diffusion de résultats 40
9.1 Animation de la communauté scientifique 40
9.2 Enseignement universitaire 40
9.3 Participation à des colloques, séminaires, invitations 41

10 Bibliographie 41

1 Composition de l'équipe

Responsable scientifique

Jacques Nicolas [CR Inria]

Assistante de projet

Maryse Auffray [AA Inria]

Personnel Inria

Yves Moinard [CR Inria]

René Quiniou [CR Inria]

Personnel Université de Rennes 1 et autres établissements d'enseignement

Catherine Belleannée [maître de conférences]

Marie-Odile Cordier [professeur, détachement Inra (depuis oct. 1998)]

Israël-César Lerman [professeur]

Véronique Masson [maître de conférences]

Laurent Miclet [professeur à l'Enssat (participe également au projet Cordial)]

Dominique Py [maître de conférences, IUFM de Rennes]

Sophie Robin [maître de conférences]

Laurence Rozé [maître de conférences, Insa de Rennes]

Pascale Sébillot [maître de conférences, délégation CNRS]

Basavaneppa Tallur [maître de conférences]

Raoul Vorc'h [maître de conférences]

Chercheurs doctorants

Laurent Blin [bourse Région]

François Coste [bourse MENRT (contrat - Ater université de Rennes 1 à partir d'octobre 1999)]

Daniel Fredouille [bourse MENRT (à partir d'octobre 1999)]

Irène Grosclaude [bourse MENRT]

Jean-Marc Guinnebault [bourse MENRT (jusqu'à février 1999)]

Christine Largouët [AERC Ensar]

Konan Lemée [bourse MENRT]

Emmanuel Mayer [bourse université (contrat - Ater université de Rennes 1 à partir d'octobre 1999)]

Yannick Pencolé [bourse MENRT]

Ronan Pichon [bourse MENRT]

Romuald Texier [bourse Cifre]

Collaborateurs extérieurs

Philippe Besnard [DR CNRS, Irit, Toulouse]
Olivier Mescam [service civil - objecteur de conscience à partir du 15 novembre 1999]
Raymond Rolland [maître de conférences, université de Rennes 1]
Valérie Rouat [Ater IUT de Lens]

2 Présentation et objectifs généraux

2.1 Présentation générale et objectifs

Notre problématique générale est de fournir une assistance intelligente à un opérateur confronté à l'analyse de données complexes et de taille importante. Il s'agit d'extraire de ces données les éléments qui permettent à l'opérateur d'agir au mieux. Ceci suppose au minimum la mise au point d'un modèle explicatif des données traitées et souvent celle d'un modèle de l'utilisateur lui-même, afin de réaliser l'interface nécessaire à cette assistance.

Par assistance intelligente, nous entendons donc le développement de capacités automatiques de modélisation, de reconnaissance de situations intéressantes et d'élaboration de recommandations d'actions adaptées et explicables.

Nous nous situons dans une perspective intelligence artificielle. Le but est de rendre l'utilisateur autonome face à l'analyse de ses données, c'est-à-dire de ne pas requérir la présence d'un tiers (spécialiste) pour l'interprétation des résultats fournis. Respecter cet objectif suppose de fournir des résultats facilement interprétables et donc de travailler sur des modèles qui restent compréhensibles par cet utilisateur.

Ce thème correspond à des besoins bien identifiés en terme d'utilisateurs : opérateur chargé de la surveillance d'un système, scientifique cherchant à découvrir des relations intéressantes dans une masse de données, utilisateur sélectionnant des documents dans une base documentaire.

Les *thèmes scientifiques* sur lesquels se focalisent le projet concernent tous des capacités fondamentales pour l'interprétation de données : il s'agit de synthèse, de généralisation ou d'abstraction. Ces capacités sont de nature essentiellement abductive (pouvoir ajouter des hypothèses pertinentes à un ensemble de connaissances pour tenir un raisonnement) ou inductive (pouvoir induire des règles à partir de connaissances de même nature), c'est-à-dire que le problème central est celui de la sélection dans un ensemble donné (généralement infini) d'une ou de plusieurs hypothèses pertinentes pouvant expliquer au mieux un ensemble d'observations. De façon plus précise, le projet s'articule en deux composantes :

- **Modélisation** de systèmes (physiques, biologiques) ou de données complexes (langage naturel), en vue du diagnostic ou plus généralement de l'extraction de l'information pertinente. On s'intéresse à des modèles symboliques, par opposition aux modèles mathématiques utilisés en automatique.

– **Apprentissage** pour l'acquisition ou la mise au point de ces modèles (essentiellement programmation logique inductive, inférence grammaticale et analyse de données). Là encore, il s'agit d'apprentissage symbolique, par opposition à des techniques d'apprentissage par renforcement.

Les *thèmes d'application* sur lesquels se focalisent le projet sont les suivants :

– **Aide à la surveillance de systèmes physiques**
Un système physique évolue dans le temps, soit du fait de sa dynamique propre, soit sous l'effet d'actions ou d'événements extérieurs. La surveillance d'un tel système consiste à analyser les observations issues de capteurs, à en inférer l'état courant du système afin de détecter un éventuel dysfonctionnement, à caractériser ce dysfonctionnement en localisant le ou les composants défectueux, et éventuellement à préconiser l'action (ou la suite d'actions) qui semble la plus appropriée au maintien ou au rétablissement des fonctionnalités du système. Nous nous limitons aux systèmes de surveillance dans lesquels un opérateur est impliqué ; il s'agit donc plus précisément d'*aide* à la surveillance d'un système.

– **Aide à l'interprétation de séquences**
Nous considérons ici deux types très différents de séquences naturelles : les textes (documents) et les séquences biologiques (ADN, ARN, protéines), vues comme des textes sur un alphabet généralement réduit. Dans les deux cas, on s'intéresse prioritairement à l'analyse de contenu. Le but est d'extraire la connaissance incluse dans les textes, en passant par une phase d'indexation automatique. Celle-ci consiste à traduire le contenu de ces textes en une structure de données facilitant la recherche lors du traitement des requêtes qui lui sont adressées. Le filtrage d'éléments pertinents nécessite de plus l'emploi d'outils d'analyse syntaxique et/ou statistique.

3 Fondements scientifiques

3.1 Aide à la surveillance de systèmes physiques

Mots clés : surveillance, diagnostic, modèle de fonctionnement, modèle de panne, simulation, reconnaissance de scénario, graphe causal temporel, acquisition de scénario.

Glossaire :

alarme indicateur discret émis par un système de surveillance à partir d'événements et censé provoquer une réaction humaine ou automatique.

scénario (ou chronique) ensemble d'événements ponctuels et de contraintes temporelles sur ces événements caractéristiques d'une situation.

reconnaissance de scénario système permettant, à partir d'un ensemble de scénarios décrivant des situations (la base de scénarios), d'analyser au vol une séquence d'observations datées et de reconnaître les situations.

Résumé :

Les principales approches de l'intelligence artificielle au problème de la surveillance (et supervision) de systèmes sont basées sur un modèle de fonctionnement ou des dysfonctionnements au cœur du système de surveillance. Nous décrivons essentiellement le domaine de la modélisation de systèmes (ou d'activités complexes) évoluant dans le temps en vue de leur surveillance en ligne. Pour plus de détails et pour les références, consulter par exemple [BC96,GRO97,GRO98].

Le problème de la supervision par gestion d'alarmes est au cœur de nos travaux. Un opérateur chargé de la surveillance reçoit des événements (les alarmes) datés et émis par les composants eux-mêmes en réaction à des événements extérieurs. Les observations recueillies sur le système sont des informations discrètes, correspondant à un événement ponctuel ou à une propriété associée à un intervalle de temps. Les principales difficultés pour analyser ce flux d'alarmes sont alors les suivantes :

- la profusion des alarmes reçues : le superviseur peut recevoir jusqu'à plusieurs centaines de messages par seconde, dont certains sont non significatifs.
- l'imbrication des alarmes reçues : les ordres dans lesquels sont émises et reçues les alarmes peuvent être différents. De plus, les séquences d'alarmes résultant de pannes concourantes peuvent s'imbriquer. Les délais de propagation et, éventuellement, les voies d'acheminement doivent ainsi être pris en compte, aussi bien pour rétablir l'ordre des événements que pour décider à partir de quand on peut supposer avoir reçu la totalité des messages pertinents.
- leur redondance : certaines alarmes sont de simples conséquences d'autres. C'est en particulier le cas dans le phénomène connu sous le nom d'avalanche d'alarmes.
- perte et masquage : certaines alarmes émises peuvent être perdues ou masquées au superviseur par suite du dysfonctionnement d'un composant intermédiaire chargé de leur transmission. L'absence d'une alarme doit être prise en compte et peut fournir une indication intéressante sur l'état du système.

On peut distinguer deux cas posant des problèmes un peu différents. Les alarmes de conduite sont destinées à être traitées *en ligne* par l'opérateur de conduite. Le but de la surveillance est alors l'aide à la conduite, et l'analyse doit être faite en temps réel. L'opérateur a un objectif d'optimisation à court terme : il s'agit en général de rester au plus près d'un régime idéal, en tenant compte de la variabilité des entrées et de l'évolution naturelle des processus. En revanche, les dérives structurelles du système (usure des pièces, modifications lentes des propriétés de ses composants, etc.) ne sont pas prises en compte en tant que telles et sont corrigées par un réglage de paramètres.

[BC96] M. Basseville, M.-O. Cordier, « Surveillance et diagnostic de systèmes dynamiques : approches complémentaires du traitement de signal et de l'intelligence artificielle », *rapport de recherche n° 1004*, IRISA, Mars 1996.

[GRO97] GROUPE ALARME, *Surveillance et interprétation d'alarmes en milieu industriel*, Actes des journées PRC-IA, Éditions Hermès, Grenoble, 1997, p. 9–30, S. Cauvin, M.-O. Cordier, C. Dousson, G. Defrandre, P. Laborie, F. Lévy, J. Montmain, M. Porcheron, I. Servet, L. Travé-Massuyès.

[GRO98] GROUPE ALARME, « Monitoring and alarm interpretation in industrial environments », *AI Communications 11, 3-4*, 1998, p. 139–173, S. Cauvin, M.-O. Cordier, C. Dousson, P. Laborie, F. Lévy, J. Montmain, M. Porcheron, I. Servet, L. Travé-Massuyès.

Ce traitement *réactif* s'oppose au traitement *en profondeur* des alarmes de maintenance. On procède, dans ce cas, à une analyse *hors ligne* plus fouillée de l'historique du système, en cherchant à prévoir les incidents, à planifier les opérations d'entretien pour limiter au maximum les défaillances et les interruptions de service.

Dans le cadre de l'aide à la surveillance, nous cherchons non seulement à détecter les dysfonctionnements mais à les caractériser en localisant les composants responsables (diagnostic). Nous utilisons les approches dites à base de modèles pour lesquelles on suppose disponibles des modèles de fonctionnement ou de dysfonctionnement des systèmes surveillés.

L'exploitation en ligne des modèles est rarement envisageable car trop complexe vis-à-vis des contraintes temps réel, ceci en particulier en raison de la dimension temporelle que ces modélisations prennent en compte (automates communicants temporels ; graphes causaux temporels). Une approche consiste à transformer ces modèles hors ligne en en extrayant les éléments utiles au diagnostic.
Deux méthodes sont étudiées :

- Dans la première, le modèle est utilisé en simulation afin d'acquérir pour chaque panne significative les séquences d'observations correspondantes et constituer ainsi une base significative d'apprentissage. Les simulations associent à chaque situation de pannes ce que l'on appelle un scénario, c'est-à-dire un ensemble d'observables et un ensemble de contraintes temporelles qu'ils doivent respecter. Une des techniques permettant la supervision de systèmes dynamiques est alors la reconnaissance à la volée de ces scénarios. Son principe consiste en un suivi, en fonction des messages reçus, d'un ensemble de scénarios potentiels jusqu'à une reconnaissance complète d'un ou plusieurs d'entre eux. L'apport d'une base de scénarios est, dans ce cas, nécessaire au bon fonctionnement de la supervision. Cette base doit contenir l'ensemble des scénarios de pannes possibles. Or son obtéention n'est pas toujours aisée. Elle doit, par ailleurs, être actualisée au fur et à mesure de l'évolution, physique ou structurelle, du système sous surveillance. Une expertise humaine régulière s'avère coûteuse, raison pour laquelle il est préférable de s'orienter vers une méthode d'acquisition automatique de scénarios. Les séquences étiquetées sont ensuite généralisées afin d'obtenir un ensemble de scénarios discriminants. Un système de reconnaissance de scénarios est alors utilisé en ligne pour la surveillance du système.
- Dans la seconde approche, l'automate qui sert de modèle est transformé hors ligne en un automate adapté au diagnostic, appelé « diagnostiqueur ». Ses transitions s'effectuent uniquement à partir des événements observables et ses états contiennent de l'information sur les pannes rencontrées par le système. Diagnostiquer le système consiste à parcourir le diagnostiqueur au fur et à mesure de l'arrivée d'événements observables.

Dans les deux cas, le point important est la réduction de la complexité. Dans la première approche, le point clé est d'extraire les informations discriminantes suffisantes pour identifier les dysfonctionnements. L'apprentissage automatique peut s'effectuer par différentes techniques. L'utilisation de la programmation logique inductive avec contraintes semble à cet égard représenter une voie de recherche intéressante. Dans le second cas, l'idée est celle de généricité et de répartition. Profitant de la structure du système (dans notre cas, la structure arborescente), le modèle générique est une représentation économique et suffisante permettant d'éviter de construire le modèle global et se contentant du modèle d'une branche.

3.2 Apprentissage automatique

Mots clés : inférence grammaticale, analyse de données, classification automatique, programmation logique inductive.

Glossaire :PTA Prefix Tree Acceptor : il s'agit du plus petit automate fini déterministe reconnaissant l'ensemble des préfixes d'un ensemble de mots donné.

programme logique ensemble fini de clauses définies.

clauses définies disjonction de littéraux contenant un seul littéral positif, un littéral étant soit une formule atomique, soit la négation d'une formule atomique.

variable au sens «analyse des données» : il s'agit d'un attribut, d'un élément d'un système descriptif

ensemble des modalités domaine, ensemble des valeurs possibles pour une variable.

Résumé : *On décrit ici les techniques étudiées dans le projet, visant à acquérir des modèles et à les mettre au point de manière automatique à partir d'un ensemble d'observations. Cette automatisation pose des problèmes de filtrage, de structuration des observations, puis de spécification du «saut inductif», c'est-à-dire de la manière dont vont être définis puis calculés les modèles acceptables au vu des observations.*

Le projet s'appuie pour cela sur les travaux issus de l'apprentissage, de la classification et de l'analyse des données. Plus précisément, nous nous intéressons à un apprentissage de type structurel, c'est-à-dire où il s'agit de faire émerger des relations entre données parmi lesquelles les dépendances ne sont pas connues. Les techniques associées ressortent de l'inférence grammaticale ou de la programmation logique inductive suivant que les structures visées sont des grammaires ou des programmes.

3.2.1 Inférence grammaticale et programmation logique inductive

On appelle inférence grammaticale l'apprentissage automatique d'un modèle de langage à partir d'un échantillon fini des phrases du langage que la grammaire accepte (instances positives) et éventuellement d'un échantillon fini de phrases n'appartenant pas à ce langage (instances négatives). Les phrases correspondent dans les applications à un ensemble d'observations sur l'état ou le comportement du système et peuvent être aussi bien des séquences biologiques, des séquences d'alarmes ou des suites d'actions.

Spécifier complètement un problème d'inférence grammaticale suppose de

- définir la classe des langages acceptés ;
- définir la représentation des langages sur laquelle on travaille (grammaires formelles, automates, expressions) ;
- définir une relation d'ordre (relation de généralité) sur ces représentations, compatible avec l'inclusion sur les langages ;
- définir les conditions de présentation des phrases d'apprentissage («oracle» répondant aux questions de l'algorithme, présentation en bloc des instances ou incrémentale) ;

- définir un critère d'acceptation des solutions en fonction des instances, qui raffine la simple acceptation des instances positives et le rejet des instances négatives dans les langages associés aux solutions ;
- enfin, spécifier une stratégie d'exploration de l'espace des représentations choisi.

Nous nous intéressons plus particulièrement aux travaux tendant à renforcer l'applicabilité pratique des techniques d'inférence. Notre objectif est de démontrer que, moyennant un certain nombre de recherches, les résultats de l'inférence grammaticale sont transférables à l'analyse de corpus réels. De façon annexe se pose le problème de la constitution de benchmarks permettant la comparaison et l'évaluation des algorithmes produits.

Nous nous restreignons au cas où la classe acceptée est la classe des langages rationnels et où on travaille sur une représentation par automates finis. Il existe une relation d'ordre de généralité naturelle sur les automates induite par la fusion d'états dans un automate : toute fusion d'états dans un automate mène à un automate (appelé automate dérivé) reconnaissant un langage plus général ou équivalent au langage reconnu initialement. Si de plus on prend comme critère d'acceptation la complétude structurelle (c'est-à-dire. toutes les transitions et états d'acceptation d'un automate sont exercés), on montre que l'espace de recherche de toutes les solutions est un treillis. Celui-ci peut être construit à partir d'un automate canonique reconnaissant uniquement les instances positives. Les éléments du treillis sont dérivés de cet élément nul (l'automate canonique) par une fonction correspondant à la fusion de ses états. L'élément universel du treillis est l'automate universel, reconnaissant n'importe quelle suite de caractères. On peut restreindre encore l'espace de recherche si l'on s'intéresse uniquement aux automates déterministes. Dans ce cas, on remplace l'automate canonique par le PTA. L'apprentissage se ramène alors fondamentalement à un problème d'énumération dans un (grand) ensemble partiellement ordonné.

Les travaux que nous développons cherchent à étendre l'applicabilité des méthodes d'inférence sur les deux points suivants, en relation avec la liste que nous avons définie précédemment :

- mode de présentation des instances : passer d'un apprentissage «à données fixées», c'est-à-dire où l'on dispose initialement de toutes les instances, à un apprentissage incrémental, où les instances peuvent être disponibles en plusieurs étapes, suppose la résolution d'un certain nombre de problèmes difficiles si on ne souhaite pas recommencer l'apprentissage à partir de zéro à chaque nouvelle instance présentée.
- stratégie d'exploration : que le critère soit explicite ou non, la plupart des méthodes se contentent de fournir une seule solution, correspondant à un minimum local. Il s'agit d'une limitation importante par rapport aux applications : la plupart du temps, l'automate ayant une vertu explicative, on souhaite une caractérisation de l'ensemble des solutions possibles (combien y en a-t-il, en quoi diffèrent-elles ?). Ceci suppose de s'attacher à l'étude de stratégies complètes.
  Un second point concerne le sens de la recherche dans l'espace des grammaires ou automates : la plupart des méthodes procèdent par fusion d'états ou de non-terminaux, suivant en cela une progression par généralité croissante. Le critère de généralité maxi-

male étant cependant souvent retenu, il est intéressant d'étudier à l'inverse l'inférence par «fission», autrement dit par spécialisation croissante d'un reconnaisseur universel. On espère ainsi aboutir aux solutions en un nombre réduit d'étapes.

La programmation logique inductive (Pli) consiste à inférer un programme logique P (par exemple, dans le langage Prolog) à partir de la donnée de faits complètement instanciés F qui doivent être vérifiés dans le programme cible et éventuellement d'un noyau de programme T qui modélise des informations déjà connues, qui peuvent faciliter l'apprentissage. Sur un plan logique, on souhaite vérifier la relation $T,P \models F$. Les prédicats pouvant intervenir dans les clauses de P sont généralement fixés, de même que l'ensemble des termes admissibles. Par rapport aux techniques d'inférence grammaticale présentées précédemment, on s'intéresse un peu au problème structurel qui consiste à trouver l'ensemble des relations intervenant dans les clauses du programme, et beaucoup au problème de la généralisation des termes intervenant dans les relations. Nous nous intéressons particulièrement aux techniques d'induction sur des clauses contraintes où les variables sont soumises à un système de contraintes [SR96]. L'étude des relations entre inférence grammaticale et programmation logique est pertinente mais reste un domaine vierge. Les résultats escomptés sont des apports croisés dans ces approches et une meilleure maîtrise de leurs domaines d'application respectifs. Un autre intérêt est de pouvoir étudier le problème de l'inférence grammaticale dans un contexte logique, où l'induction est ramenée à un problème d'unification.

3.2.2 Classification

La classification est l'étape la plus en amont d'un processus d'analyse, étape considérant les données de manière globale, qui va faciliter des analyses postérieures plus fines, en regroupant ou au contraire en discriminant des ensembles de données brutes. L'enjeu et l'objectif est donc celui de la réduction la plus importante de la complexité qui permette cependant de filtrer au mieux l'information significative. Le contexte général où se situent nos travaux est celui d'une interaction entre d'une part, une approche de classification non métrique, combinatoire et statistique et, d'autre part, un ensemble de problèmes algorithmiques fondamentaux qui se présentent dans l'analyse de données complexes issues de l'observation, de la connaissance ou de modèles.

L'aspect classification comprend aussi bien la classification non supervisée par Analyse de la Vraisemblance du Lien (AVL) que celle supervisée qui relève de la discrimination par arbres de décision. D'autres méthodes d'analyse combinatoire des données peuvent également intervenir.

La classification est un outil fondamental pour spécifier une algorithmique de résolution approchée ; inversement, l'algorithmique intervient de façon essentielle dans la résolution de nos problèmes combinatoires de classification.

Les thèmes scientifiques que nous développons concernent les points suivants :

– réduction de la complexité d'un système descriptif (e.g. classification pour l'inférence de connaissances lexicales à partir de corpus de textes, voir la section suivante) ;

[SR96] M. Sebag, C. Rouveirol, « Induction de clauses contraintes », *in : Reconnaissance des formes et intelligence artificielle (RFIA'96)*, p. 706–716, Rennes, 1996.

- élaboration de coefficients d'associations (e.g. pour la classification de parcelles agricoles, à partir d'images en vue de la surveillance d'une zone) ;
- comparaison de classifications sur des données complexes (e.g. pour la comparaison de différentes classifications de parcelles agricoles obtenues avec différents paramètres).

3.3 Recherche d'information dans un ensemble de documents, construction de lexiques

Mots clés : recherche d'information, terminologie, séquence binominale, calcul sémantique, acquisition d'informations lexicales en corpus.

Glossaire :

composé, séquence binominale, structure binominale complexe dans nos travaux, association de deux noms de la forme *Nom Nom* en anglais et *Nom à/de (déterminant) Nom* en français. Ces noms peuvent être simples ou obtenus par adjonction d'un suffixe à un verbe (constituant déverbal).

interprétation ou calcul sémantique d'un composé détermination de la relation qu'entretiennent les constituants d'un composé.

prédicat, arguments un prédicat désigne un opérateur mettant en relation des arguments. Dans la phrase, le verbe joue en général le rôle de prédicat, les compléments étant ses arguments. La liste des arguments d'un prédicat forme sa structure argumentale.

terme symbole conventionnel qui désigne de façon univoque une notion à l'intérieur d'un domaine de connaissances.

Résumé : *Nous nous intéressons à la modélisation du contenu des textes via la modélisation de la sémantique de ses éléments descripteurs en indexation automatique. Notre but est de fournir des méthodes linguistiques permettant d'augmenter les possibilités d'apparier une requête et les textes de la base documentaire. Nous proposons d'une part un modèle hors domaine dont la fonction est de calculer le sens des séquences complexes*[1]*, qui constituent l'essentiel des termes des domaines techniques, en rétablissant leurs structures prédicatives sous-jacentes, et nous acquérons d'autre part les informations lexicales nécessaires à ce calcul (et à son extension en domaine) de manière automatique sur corpus. Ce module présente les idées centrales des différents thèmes que nous abordons, la recherche d'information, l'analyse de séquences complexes (extraction et interprétation) et l'acquisition d'informations en corpus.*

3.3.1 Recherche d'information - Indexation automatique

La recherche d'information (recherche documentaire) consiste, à partir d'un ensemble de textes et d'une requête d'un utilisateur, à proposer à ce dernier les textes adéquats. Il convient donc d'identifier les notions importantes d'un texte et de mesurer la proximité entre une requête et les textes de la base en déterminant celles qu'ils partagent. Les travaux de ce domaine

1. Nous utilisons ici indifféremment composé, séquence complexe, ou séquence binominale dans le cas de deux noms.

passent généralement par une phase d'indexation automatique[SM83]. La qualité des systèmes de recherche d'information dépend de ce fait largement des techniques employées pour traduire le contenu des textes dans un langage d'indexation et pour réaliser l'appariement entre les textes indexés de la base consultée et la requête. Leur performance est mesurée à l'aide du *rappel*, proportion de réponses retrouvées parmi celles à produire, et de la *précision*, proportion de réponses pertinentes retrouvées parmi celles produites.

On oppose en général deux types d'indexation : l'indexation par index atomiques (indexation simple), qui assimile les indicateurs de contenu aux mots simples du texte (objectif premier : le rappel) mais conduit à des index peu discriminants et ambigus, et l'indexation par index complexes (indexation syntagmatique), qui manipule des groupes de mots (objectif premier : la précision) et aboutit donc à des index plus spécifiques et plus dispersés. En fait, les résultats des systèmes ayant choisi l'une ou l'autre option ne permettent pas de trancher de manière définitive entre ces deux techniques et une voie moyenne semble raisonnable. Une façon d'aboutir à ce résultat consiste à privilégier une indexation syntagmatique sémantiquement riche, afin d'augmenter les possibilités d'appariement entre une requête et les textes de la base documentaire.

3.3.2 Analyse des séquences complexes

L'analyse des séquences complexes, en particulier binominales, est un enjeu fondamental dans de nombreuses applications du traitement automatique du langage naturel (TALN).

Une première phase de cette analyse, qui a fait l'objet de nombreux travaux, concerne l'extraction automatique de ces séquences qui constituent une grande proportion des termes, surtout dans les domaines scientifiques. Le repérage des séquences candidates à être des termes s'effectue selon les systèmes, soit par des critères syntaxiques, soit par des critères essentiellement statistiques, soit par une approche mêlant ces deux aspects (cf. par exemple[Bou94,Dai94]).

Une seconde direction concerne l'analyse sémantique de ces séquences. L'objectif des travaux de ce domaine est fréquemment de trouver la relation prédicative qui lie les constituants des composés. La difficulté du problème abordé tient au fait qu'une part importante de l'information sémantique contenue dans les séquences composées est implicite, ce qui nécessite de rendre compte d'inférences complexes. Par exemple, un *interpréteur de commandes* sert à *interpréter* des commandes (*relation explicite*) alors qu'un *parc à munitions* sert à *entreposer* des munitions (*relation implicite*). Le caractère implicite est de plus source d'ambiguïtés : *milk disease* est une maladie *causée* par le lait alors que *plant disease* est une maladie *affectant* une plante. De très nombreux travaux ont été consacrés, tant en linguistique qu'en intelligence artificielle, à la question de la détermination automatique du sens des séquences complexes à partir de la représentation sémantique des éléments simples qui les composent. Dans le domaine du TALN, deux types de modèles s'opposent : ceux qui dépendent d'un domaine, et ceux qui se consacrent à l'interprétation hors domaine des composés. C'est dans ce dernier cadre que

[SM83] G. Salton, M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Computer Science Series, 1983.

[Bou94] D. Bourigault, *Acquisition de terminologie*, thèse de doctorat, EHESS, 1994.

[Dai94] B. Daille, *Approche mixte pour l'extraction de terminologie : statistiques lexicales et filtres linguistiques*, thèse de doctorat, Université Paris 7, 1994.

se situent nos travaux. Les systèmes hors domaine proposent diverses stratégies. Une d'entre elles consiste à fonder le calcul de la sémantique des séquences complexes sur des règles générales d'interprétation, qui associent des prédicats à certains noms simples (c'est-à-dire non déverbaux) et font jouer à l'autre constituant de la séquence un rôle dans la structure argumentale de ce prédicat. Cette approche, initialisée par Finin[Fin80], trouve des prolongements dans les travaux développés au sein de notre équipe, où un modèle général d'interprétation hors domaine des composés, basé sur les travaux de Lieber[Lie83] et Selkirk [Sel82] pour traiter les composés déverbaux, et sur le modèle du Lexique Génératif de Pustejovsky[Pus95] pour interpréter les séquences complexes à relation implicite, a été développé.

Quelle que soit la méthode utilisée pour définir des mécanismes de calcul de la sémantique des séquences composées, l'interprétation passe par l'étude précise de la sémantique nominale. Les lexiques correspondants ne peuvent pas être construits manuellement pour chaque application et ces informations lexicales doivent donc être acquises automatiquement à partir de corpus de textes du domaine de l'application visée.

3.3.3 Acquisition automatique d'informations lexicales à partir de corpus

Le développement de travaux d'acquisition automatique d'informations lexicales à partir de corpus connaît un essor considérable depuis le début des années 90 (cf.[11] pour un bilan du domaine).

Outre les travaux en extraction de terminologie présentés plus haut, l'acquisition consiste principalement à rechercher par des techniques statistiques les informations sur les unités extraites. Celles-ci sont de deux types : syntagmatiques et paradigmatiques.

Les informations syntagmatiques concernent les capacités d'association d'un mot : étant donné un mot, on cherche à découvrir les mots qui apparaissent dans le même contexte. Les travaux de ce type s'intéressent par exemple à trouver la structure argumentale de prédicats, à repérer des verbes typiquement associés à des noms, etc. Les informations paradigmatiques concernent les similarités entre les mots : étant donné un mot, on cherche à découvrir les mots qui ont des comportement les plus proches, c'est-à-dire, en se basant sur les thèses de Harris[HGR+89], ceux qui génèrent les mêmes contextes. Les travaux de ce type cherchent par exemple à constituer automatiquement des classes sémantiques ou à découvrir des relations lexicales (synonymie, antonymie, etc.) entre des mots.

[Fin80] T. Finin, *The Semantic Interpretation of Compound Nominals*, thèse de doctorat, University of Illinois, 1980.

[Lie83] R. Lieber, « Argument Linking and Compounds in English », *Linguistic Inquiry 2*, 14, 1983, p. 251–285.

[Sel82] E. Selkirk, « The Syntax of Words », *MIT Press*, 1982.

[Pus95] J. Pustejovsky, *The Generative Lexicon*, Cambridge:MIT Press, 1995.

[HGR+89] Z. Harris, M. Gottfried, T. Ryckman, P. M. Jr, A. Daladier, T. Harris, S. Harris, « The Form of Information in Science, Analysis of Immunology Sublanguage », *Boston Studies in the Philosophy of Science 104*, 1989.

4 Domaines d'applications

4.1 Panorama

Résumé : *Les principaux domaines d'application des travaux de recherche menés dans le projet sont la génomique, la supervision de réseaux de télécommunication et la recherche d'information. Plus récemment le «monitoring» de l'activité cardiaque ainsi que la surveillance dans le domaine de l'environnement : transfert de polluants tels que pesticides et nitrates, surveillance de l'évolution des parcelles agricoles. D'autres applications sont abordées dans des domaines connexes tels que l'étude des séquences de mots en reconnaissance de la parole.*

4.2 La génomique

Mots clés : automate, bio-informatique, analyse linguistique.

Résumé : *L'objectif est, à partir d'un ensemble d'observations de suites produites par un mécanisme mal connu (des macro-molécules biologiques), de mettre en évidence des sous-séquences ou des structures qui apportent des indices pour la compréhension de ce mécanisme.*

L'objectif est, à partir d'un ensemble d'observations de suites produites par un mécanisme mal connu (des macro-molécules biologiques), de mettre en évidence des sous-séquences ou des structures qui apportent des indices pour la compréhension de ce mécanisme.

Ceci peut s'effectuer par la recherche des sous-séquences «surprenantes». En effet, l'existence de «signaux» biologiques se repère dans une séquence génétique par des sous-séquences particulières anormalement répétées ou enchaînées de façon précise. Nous avons ainsi étudié le mécanisme d'initiation de la traduction chez *E. coli* et nous proposons de même d'étudier l'enchaînement de motifs particuliers tout au long du génome. Le but peut être également de modéliser un mécanisme particulier en établissant une correspondance entre séquence, structure et fonction. Ainsi, nous avons commencé à étudier le phénomène de régulation dans les gènes impliqués dans la lipogénèse sur les vertébrés, qui fait intervenir des motifs encore peu connus, de taille très réduite et donc difficiles à repérer individuellement mais dont la structure d'enchaînement est relativement précise (e.g. palindromes faiblement espacés). Bien que le domaine des séquences biologiques soit un domaine d'intérêt privilégié, la classe d'applications permet d'envisager des domaines très variés où les mêmes techniques sont utilisables. Nous avons ainsi un contrat Cnet en cours sur l'inférence de la syntaxe en reconnaissance de la parole et d'autres projets de recherches possibles en collaboration avec des industriels (modélisation de la stratégie d'un apprenant dans la résolution d'un problème par étapes ou dans son parcours d'un logiciel d'enseignement, automate d'accès à un service à partir de séquences).

Les difficultés peuvent provenir de la taille des séquences, de l'existence d'interactions à longue distance, et de la superposition de nombreuses contraintes indépendantes pour aboutir à la séquence observée. Comme dans tout domaine réel, il faut aussi résoudre des problèmes d'approximation ou de bruit sur les observations. La modélisation s'attache à décrire les séquences à un niveau lexical, syntaxique et éventuellement sémantique.

4.3 Surveillance de systèmes physiques

Mots clés : surveillance, diagnostic, reconnaissance et acquisition de scénarios, diagnostiqueur, réseaux de télécommunications, surveillance cardiaque, systèmes natures, parcelles agricoles.

Résumé :

L'objectif est de surveiller à partir d'un ensemble d'observations (souvent datées ou au moins ordonnées) le fonctionnement d'un système physique, de détecter un éventuel dysfonctionnement, de le caractériser en localisant le ou les composants défectueux et de préconiser l'action (ou la suite d'actions) qui semble la plus appropriée au maintien des fonctionnalités du système. Nous nous appuyons sur les méthodes utilisant des modèles de fonctionnement ou de dysfonctionnement des systèmes surveillés (approches de type model-based) tout en cherchant à construire des systèmes efficaces utilisables en temps réel.

L'objectif est de surveiller à partir d'un ensemble d'observations (souvent datées ou au moins ordonnées) le fonctionnement d'un système physique, de détecter un éventuel dysfonctionnement, de le caractériser en localisant le ou les composants défectueux et de préconiser l'action (ou la suite d'actions) qui semble la plus appropriée au maintien des fonctionnalités du système.

Les systèmes auxquels nous nous intéressons étant pour la plupart dynamiques (évolution dans le temps), les modélisations sur lesquelles nous nous focalisons permettent de tenir compte de la dimension temporelle : automates communicants temporels, graphes causaux temporels, logiques du changement et logiques de l'action. Nous appliquons nos méthodes à la surveillance de systèmes physiques aussi bien artefacts (tels que les réseaux de télécommunications) que naturels (tels que le système cardiaque ou les systèmes écologiques) :

- **Surveillance de réseaux de télécommunications.** Deux types d'approches sont expérimentées pour la surveillance de ces réseaux. La première approche est de type reconnaissance de scénarios et tire parti de l'efficacité de ce type de méthodes pour satisfaire aux contraintes temps réel. Un des points importants est l'acquisition automatique de ces scénarios afin en particulier de pouvoir prendre en compte l'évolution technologique rapide des systèmes considérés. Nous privilégions une approche de type apprentissage supervisé en nous appuyant sur les modèles décrivant leur fonctionnement. L'acquisition des scénarios se fait à partir des données résultant de la simulation de dysfonctionnements et fait appel à des techniques d'apprentissage de type Pli (programmation logique inductive et, plus particulièrement, Pli avec contraintes). Nous appliquons cette approche à la surveillance de réseaux de télécommunications dans le cadre d'une collaboration avec le Cnet dans le cadre du projet Gaspar (contrat de type CTI) et du projet Magda (contrat RNRT). Une autre approche consiste à *compiler* le modèle du système, représenté par un graphe causal temporel, en un ensemble de scénarios. Le modèle est utilisé de manière déductive et l'interaction entre pannes multiples est prise en compte. Cette approche est appliquée au diagnostic de pompes primaires dans le cadre d'un contrat avec l'EDF.

La seconde approche consiste à produire directement un automate diagnostiqueur à partir de l'automate modélisant le comportement du système. Les travaux actuels ont pour objectif la construction de diagnostiqueurs génériques (pour ne pas avoir à représenter l'ensemble des comportements instanciés de chacun des composants mais uniquement leurs classes de comportement), ainsi que de diagnostiqueurs décentralisés (afin de pouvoir répartir une partie du diagnostic au niveau des composants eux-mêmes en s'appuyant sur des diagnostiqueurs locaux). Cette approche est appliquée aux réseaux de télécommunications au sein du projet Magda (contrat RNRT).

- **Surveillance cardiaque.** La technique de reconnaissance de scénarios est utilisée pour la surveillance, à partir de leur électrocardiogramme, de patients souffrant de problèmes cardiaques. Les scénarios sont obtenus par apprentissage automatique (Pli) sur des données provenant de simulations et de signaux réels. Nous prévoyons d'étendre la méthode dans le cadre de la conception d'une prothèse cardiaque (pacemaker - défibrillateur) «intelligente» afin d'analyser plus finement les dysfonctionnements constatés et de produire une stimulation mieux située dans le cycle cardiaque.
- **Surveillance de systèmes naturels.** Une première application porte sur la surveillance de parcelles agricoles et s'appuie sur une suite d'images satellitales et aériennes. Après une étape de classification de ces images (classification des parcelles), les résultats sont améliorés en tirant parti de modèles de l'évolution de la couverture de ces zones agricoles. Ces modèles d'évolution sont décrits dans le formalisme des automates temporels et utilisent plus précisément le formalisme de Kronos [Yov97]. Les résultats obtenus montrent une amélioration notable dans la précision des identifications des parcelles traitées.

  Nous avons aussi abordé dans le cadre d'une collaboration avec l'Inra (Unité Sciences du Sol et Agronomie de Rennes-Quimper) deux études portant sur la modélisation du transfert du nitrate au niveau d'un bassin versant d'une part, et de pesticides au niveau d'une parcelle agricole d'autre part. Dans les deux cas, nous avons choisi de nous appuyer sur les modèles quantitatifs classiquement utilisés afin de construire des modèles qualitatifs, plus adaptés à une prise de décision. Deux prototypes ont été construits et sont en cours de validation.

4.4 La recherche d'information et l'accès à des bases de documents ou de services

Mots clés : recherche d'information, sémantique lexicale.

Résumé : *La recherche d'information constitue le domaine global d'application de nos travaux. Nous avons intégré certains de nos résultats sur le calcul sémantique des composés à un serveur d'interrogation de services télématiques dans le cadre d'un contrat* CTI *avec le Cnet. Deux voies ont été explorées : l'utilisation du contexte de la séquence binominale pour mettre en œuvre une désambiguïsation plus*

[Yov97] S. YOVINE, « Kronos: A verification tool for real-time systems », *International Journal of Software Tools for Technology Transfer 1*, 1997.

efficace des mots et le réordonnancement des réponses proposées en favorisant celles obtenues en suivant les liens de modification nominale.

L'amélioration de la qualité de service d'une application passe par l'adaptation de l'interaction au comportement de l'utilisateur. Notre approche consiste à interpréter les actions de cet utilisateur de façon à suivre l'évolution de ses buts et ses intentions. Ces connaissances sont modélisées par une logique modale.

4.4.1 Recherche d'information

Nous explorons trois voies complémentaires pour améliorer les performances des systèmes : le développement d'un modèle d'interprétation hors domaine des séquences binominales, l'étude de la variation sémantique des termes, c'est-à-dire la reconnaissance de l'équivalence conceptuelle de deux structures différentes, et l'inférence de connaissances lexicales à partir de corpus pour obtenir des lexiques sémantiques nécessaires au fonctionnement du modèle d'interprétation.

Une première application concrète des méthodes développées a été faite dans le cadre d'un contrat CTI avec le Cnet, dans laquelle nous avons intégré certains résultats sur le calcul sémantique des composés à un serveur d'interrogation de services télématiques. Compte tenu des contraintes du système, deux voies ont été explorées : l'utilisation du contexte de la séquence binominale pour mettre en œuvre une désambiguïsation plus efficace des mots et pour rechercher des liens de paraphrase sémantique entre la requête adressée au système et les services de la base indexée, et le réordonnancement des réponses proposées aux utilisateurs en favorisant celles obtenues en suivant les liens sémantiques de nature syntagmatique (liens de modification nominale) qui unissent les constituants des séquences complexes. Cette application a donné lieu à la publication [24].

Toujours dans cette optique, nous travaillons actuellement, dans le cadre d'une Action de Recherche Partagée de l'AUPELF-UREF en collaboration avec Pierrette Bouillon (Issco Genève), Laurence Jacqmin (Université libre de Bruxelles) et Cécile Fabre (ERSS Toulouse) à un projet dont l'objectif est de déterminer, grâce à l'utilisation de méthodes d'apprentissage sur des corpus étiquetés, les prédicats témoins des différentes facettes sémantiques des noms d'un domaine, telles qu'elles sont définies dans le lexique génératif de Pustejovsky, et de montrer l'apport de l'usage d'un tel lexique dans un système de recherche d'information.

Le travail que nous avons réalisé sur le calcul sémantique des séquences complexes peut trouver des applications dans des domaines autres que la recherche d'information, tels que la structuration de données terminologiques, le résumé automatique de textes ou la traduction automatique.

4.4.2 Système coopératif d'accès à un ensemble de services

Il s'agit d'interpréter les actions d'un utilisateur d'un service automatisé de façon coopérative, en tenant compte de l'évolution des buts et des intentions de cet utilisateur. Ce travail trouve une application particulière, en coopération avec le Cnet/France-Télécom, au sein d'un service d'interrogation oral avec un service d'informations. Les séquences à interpréter sont alors l'historique du dialogue. Une logique modale complexe, combinant divers systèmes mo-

daux, dont certains très classiques comme KD45, est déjà utilisée comme langage de représentation des connaissances. L'historique du dialogue est ainsi traduit sous la forme d'une formule modale de plus en plus volumineuse, représentant l'état de croyance actuel du système, lequel conserve ainsi également la mémoire de ses croyances passées, à chaque stade du dialogue. Il convient de tenir compte d'erreurs toujours possibles, soit parce que la requête de l'utilisateur est effectivement erronée (*donnez moi le serveur de météo marine pour l'Orne*, par exemple), soit par la suite d'une erreur du système «en amont» (de reconnaissance vocale par exemple). Il faut aussi tenir compte de l'évolution possible de la requête de l'utilisateur, qui réagit en fonction des réponses que le système lui a déjà données.

5 Logiciels

5.1 Unam : logiciel d'apprentissage inductif avec contraintes

Participant : Emmanuel Mayer.

Mots clés : programmation logique inductive ; programmation logique avec contraintes ; biais de langage ; opérateurs de raffinement.

Résumé : *Unam est un logiciel d'apprentissage automatique utilisant la programmation logique inductive (Pli) et permettant l'apprentissage de clauses avec contraintes.*

Disposant d'un ensemble d'exemples positifs et négatifs, il s'agit de trouver un concept général expliquant les exemples positifs sans toutefois appréhender les notions découlant des exemples négatifs. Cet algorithme de discrimination repose sur le parcours d'un espace d'hypothèses préalablement défini par l'utilisateur. La particularité du langage de description des hypothèses d'Unam est d'offrir la possibilité d'exprimer des contraintes de nature quelconque.

Unam a été développé dans le cadre d'un contrat avec le Cnet sur le thème de l'apprentissage de scénarios pour la supervision de réseaux de télécommunications [14]. Il a servi à l'apprentissage de scénarios à partir de séquences d'alarmes étiquetées par un type de panne. Grâce à l'introduction de contraintes, notamment d'ordre temporel, dans le langage de biais, nous avons obtenu des scénarios concis et intelligibles, caractéristiques d'une panne particulière.

Le prototype est écrit en Prolog (Sicstus 3.7) et est disponible auprès de l'auteur.

6 Résultats nouveaux

6.1 Modélisation de systèmes (ou d'activités complexes) évoluant dans le temps en vue de leur surveillance en ligne

Mots clés : modélisation, supervision, diagnostic, acquisition de scénarios, apprentissage par Pli, décision en univers incertain.

Résumé : *Dans le cadre de l'aide à la surveillance de systèmes ou d'activités complexes, nous nous intéressons plus spécifiquement au cas de la surveillance par analyse de séquences d'alarmes reçues par l'opérateur. Nous cherchons non seulement à détecter les dysfonctionnements mais à les caractériser en localisant les composants responsables (diagnostic). Nous utilisons pour cela des modèles du système, en particulier des modèles de pannes, qui sont décrits dans le formalisme des automates communicants temporels pour les deux applications de surveillance des réseaux (télécommunications et distribution d'électricité) que nous traitons, ainsi que dans celui des graphes causaux temporels.*

Les activités du projet dans ce thème portent sur trois points : l'acquisition de scénarios à partir de modèles, la construction d'automates diagnostiqueurs et l'interaction diagnostic/décision dans un univers incertain. Ces travaux de recherche s'appuient principalement à la surveillance de réseaux de télécommunication dans le cadre d'un contrat CTI *avec le Cnet/France-Télécom (en collaboration avec l'université Paris-Nord) et la participation au projet* RNRT *Magda (en collaboration avec les projets Sigma2 et Pampa, l'université Paris-Nord ainsi qu'Ilog et Alcatel. Une autre application est en cours avec l'Ensar pour la surveillance de terrains agricoles à partir d'images satellitales.*

6.1.1 Acquisition de scénarios à partir de modèles

Acquisition automatique de scénarios

Participants : Marie-Odile Cordier, Emmanuel Mayer.

Les scénarios peuvent être vus comme le résultat d'une discrimination sur un ensemble de séquences de messages, étiquetées par la panne d'origine. Chaque séquence traduit les conséquences, visibles au niveau du centre de supervision, liées à une panne. L'idée est d'exhiber les caractères communs aux séquences de messages d'une même panne, et discriminants vis-à-vis des séquences de messages des autres pannes. La méthode de discrimination mise en œuvre se fonde sur les techniques de programmation logique inductive avec contraintes. Ces travaux sont décrits dans [14].

L'acquisition automatique de scénarios à partir de séquences étiquetées de messages constitue une activité de recherche bien spécifique dans le domaine de l'apprentissage. En effet, deux caractéristiques essentielles apparaissent dans les séquences de messages : leur datation et la présence multiple d'un même message au sein d'une séquence.

L'occurence multiple de messages nous a amenés à définir un nouveau formalisme et à étendre des algorithmes développés dans le cadre de la programmation logique inductive (Pli). Par ailleurs, la pertinence des contraintes liant les messages au sein d'une séquence nous a incités à introduire des termes contraints dans notre espace de recherche. La réalisation pratique de ces travaux s'appuie sur un langage de programmation classique doté d'un gestionnaire de contraintes.

Afin de valider cette approche, un modèle de fonctionnement du réseau Transpac a été réalisé sous le logiciel Asa+[2], logiciel dédié à la modélisation et à la simulation d'automates

2. logiciel commercialisé par Verilog

temporels. La base construite de séquences comporte 400 séquences classées selon cinq pannes. Les scénarios appris atteignent une qualité de discrimination supérieure à 98%. L'objectif à court terme est de les intégrer dans un logiciel de reconnaissance en ligne de scénarios du type CRS[3].

Graphes causaux temporels

Participants : Marie-Odile Cordier, Irène Grosclaude, René Quiniou.

Nous étudions une utilisation déductive des graphes causaux temporels qui, au contraire de l'approche abductive, descend intuitivement les chaînes causales pour recueillir tous les observables impliqués par la supposition de l'existence d'une (ou plusieurs) panne(s). Cette méthode permet la compilation du graphe causal en un ensemble de *scénarios*.

Le problème a été abordé en supposant l'absence d'effets contraires ou additifs dans le graphe causal. Nous avons mis en œuvre une méthode récursive permettant d'obtenir le scénario d'une panne à partir des scénarios de ses effets directs. Cette approche met l'accent sur l'efficacité du calcul des contraintes temporelles entre les observables contenus dans les scénarios. La méthode a été testée sur un graphe causal temporel représentant les dysfonctionnements possibles du circuit de refroidissement d'une centrale nucléaire. La comparaison des scénarios obtenus à partir du graphe causal avec les scénarios proposés par les experts EDF a permis de valider notre méthode [35]. Cette application a aussi mis en avant l'intérêt de la méthode dans le cadre de l'acquisition et de la validation des modèles : la comparaison des scénarios produits par compilation avec ceux donnés explicitement par les experts permet de détecter d'éventuelles incohérences dans le modèle causal. Dans une phase d'acquisition des modèles, les différences détectées peuvent servir à orienter la correction des modèles.

Même en supposant l'absence d'effets contraires ou additifs dans le graphe causal, des interactions sont possibles entre les effets de plusieurs pannes. Elles correspondent à des phénomènes de recouvrements temporels d'occurrences d'effets identiques. Ces recouvrements peuvent conduire à des observations anormales pendant une durée plus longue que celle correspondant à la superposition des durées provoquées isolément par chaque panne. Il est possible de détecter ces phénomènes en étudiant les contraintes temporelles dans le graphe causal. Le calcul de tous les scénarios possibles (correspondants à tous les recouvrements temporels possibles de causes) se heurte à une explosion combinatoire. Pour l'éviter, nous proposons de calculer un unique scénario, englobant tous les autres.

L'hypothèse de l'absence d'effets opposés ou additifs dans le graphe causal ne correspond pas à la réalité. Nous orientons actuellement notre travail vers l'étude de telles interactions, dont le résultat peut être des masquages d'effets, des modifications de la durée des effets, ou encore la production d'effets supplémentaires.

Monitoring en cardiologie

Participants : Marie-Odile Cordier, René Quiniou.

Nous étudions, en collaboration avec le LTSI (unité Inserm, université de Rennes 1), l'application en cardiologie de la surveillance par reconnaissance de scénarios. Il s'agit d'analyser

3. Logiciel distribué par le Cnet.

le signal provenant des différentes voies d'un monitoring cardiaque afin d'y détecter et de caractériser les arythmies cardiaques d'un patient sous surveillance. La nature, les caractéristiques et la fréquence des arythmies détectées permettent ensuite de proposer une attitude thérapeutique adaptée, par exemple un traitement médicamenteux ou la pose d'un pacemaker.

Une arythmie cardiaque peut se caractériser sur l'électrocardiogramme (ECG) par la succession d'ondes P et QRS respectant un certain nombre de contraintes temporelles. Il est donc naturel d'associer une arythmie à un scénario. Notre objectif est d'élaborer un outil qui puisse être adapté, autant que possible, au patient sous surveillance. En particulier, étant donnés les antécédents du patient, il n'est pas nécessaire de chercher à déceler toutes les arythmies cardiaques possibles, mais uniquement celles que le patient est susceptible de développer. La définition des scénarios à reconnaître peut alors être optimisée pour reconnaître de manière efficace le sous-ensemble d'arythmies correspondant. Nous utilisons un outil d'apprentissage automatique pour produire les scénarios à partir d'exemples d'ECG représentatifs des arythmies. Comme les scénarios sont constitués de relations temporelles entre des événements, nous avons utilisé ICL, [RL95], un système d'apprentissage automatique basé sur la programmation logique inductive (Pli) qui produit des représentations du premier ordre.

Les premiers résultats [23] obtenus sur une base d'exemples provenant d'un simulateur cardiaque sont très encourageants. Leur forme déclarative permet une validation immédiate par un cardiologue. De plus, la méthode d'apprentissage permet de paramétrer la quantité d'information utilisée pour décrire un scénario. En augmentant le nombre de cycles cardiaques devant apparaître nécessairement dans les descriptions, on obtient des scénarios plus compréhensibles, éventuellement plus robustes au bruit. En minimisant la quantité d'information nécessaire, on obtient des scénarios discriminants plus compacts et donc plus faciles à reconnaître.

Ces résultats vont être étendus dans le cadre de l'Action Concertée Incitative *Télémédecine et Technologies pour la Santé* du MENRT (cf. 7.8).

6.1.2 Construction d'un automate diagnostiqueur

Participants : Marie-Odile Cordier, Yannick Pencolé, Sophie Robin, Laurence Rozé.

La méthode des automates diagnostiqueurs s'inspire des travaux de [SSL+95,SSL+94]. Partant d'un modèle de fonctionnement d'un système décrit en terme d'automates, elle consiste à construire directement un automate particulier appelé *diagnostiqueur*. Les transitions de cet automate correspondent aux événements observables et ses états décrivent les pannes du système. Diagnostiquer le système consiste à parcourir le diagnostiqueur au fur et à mesure de l'arrivée d'événements observables. Nous avons étendu cette méthode pour l'appliquer à l'interprétation d'alarmes de réseaux de télécommunications. Le réseau est modélisé en utilisant

[RL95] L. D. Raedt, W. V. Laer, « Inductive Constraint Logic », *in : Proceedings of the 5th Workshop on Algorithmic Learning Theory, Lecture Notes in Artificial Intelligence*, 1995.

[SSL+95] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, D. Teneketzis, « Diagnosability of discrete event systems », *in : Proceedings of the International Conference on Analysis and Optimization of Systems, 40*, p. 1555–1575, 1995.

[SSL+94] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, D. Teneketzis, « A discrete event systems approach to failure diagnosis », *in : Proceedings of the Fifth international workshop on principles of diagnosis (DX'94)*, p. 269–277, 1994.

le formalisme des automates communicants temporels. L'approche diagnostiqueur développée par Sampath et al. a dû ainsi être adaptée et étendue à ce formalisme ainsi qu'aux exigences de l'application traitée.

Un prototype de ce diagnostiqueur, baptisé *Dyp*, a été réalisé. Il permet de

- construire un modèle global du système à partir d'une description en termes de composants élémentaires et d'interconnexions ;
- construire l'automate diagnostiqueur à partir du modèle global du système ;
- visionner les diagnostics réalisés au fur et à mesure de l'arrivée d'alarmes.

Le prototype précédent ne traite pas pour l'instant les informations temporelles. Il s'appuie sur un modèle global du système. Or pour un réseau de télécommunications, un tel modèle posséderait beaucoup trop d'états. C'est pourquoi nous avons étudié et implémenté des algorithmes dits génériques, profitant de la structure hiérarchique du réseau étudié. Là aussi, un prototype a été développé ; celui-ci s'appuie sur *Dyp* et permet de visualiser des groupes de branches du réseau ayant le même comportement ainsi que l'évolution des groupes au fur et à mesure de l'arrivée d'alarmes.

La construction de diagnostiqueurs décentralisés constitue actuellement un autre sujet d'étude. Le principe est de construire un diagnostiqueur local pour chaque composant du réseau et de faire coopérer ces diagnostiqueurs locaux afin d'en inférer l'état global (ou diagnostic global) du système. Nous avons ainsi adapté le diagnostiqueur afin qu'il puisse non seulement diagnostiquer les pannes du composant lui-même mais aussi tenir compte des interactions avec les autres composants du système.

Ces travaux s'effectuent dans le cadre de la poursuite du contrat CTI avec le Cnet (Projet Gaspar) ainsi que dans le cadre du contrat RNRT qui rassemble sous le nom de Magda des industriels et des équipes de recherche sur le thème de l'interprétation des alarmes dans les réseaux de télécommunications (voir section 7.1).

6.1.3 Surveillance de parcelles agricoles

Participants : Marie-Odile Cordier, Christine Largouët.

Dans le cadre d'une collaboration avec l'Ensar, nous avons abordé le problème de la surveillance de parcelles agricoles à partir d'une série d'images aériennes et satellitales avec pour objectif la maîtrise de la qualité de l'eau. Le site de l'étude est le bassin versant Chèze-Canut, d'une surface de 8000 hectares, qui alimente en eau la ville de Rennes. L'objectif du projet est de fournir, trois fois par an, une carte thématique qui résume les différentes occupations du sol (prairie, maïs, blé, etc.) des parcelles agricoles de cette région.

La classification des images par des méthodes statistiques traditionnelles (maximum de vraisemblance, nuées dynamiques, analyse discriminante) donne des résultats globalement corrects mais comportant néanmoins des anomalies ou des incohérences apparaissant sur la carte thématique résultat. Les anomalies correspondent à la dispersion de pixels isolés d'une certaine culture dans une parcelle connue comme appartenant à une autre culture. Les incohérences, détectables, si l'on compare plusieurs cartes résultats à des dates différentes, ont pour origine l'ambiguïté possible entre deux ou plusieurs cultures ayant des signatures spectrales proches.

Partant de ce constat, notre objectif est de proposer une méthode d'interprétation d'un territoire agricole par classification «intelligente» sur une séquence d'images. Nous orientons notre démarche selon deux axes : une préclassification sur la parcelle et non plus sur le pixel et la discrimination des occupations du sol à l'aide d'un modèle d'évolution de la parcelle. La préclassification a pour objectif de fournir les occupations du sol possibles pour chaque parcelle. Cette préclassification est ensuite précisée à l'aide des connaissances sur les cycles culturaux et de l'historique des observations.

La préclassification est réalisée à l'aide du logiciel Arkémie (développé par la société Arkémie Toulouse) qui propose une méthode simple de classification par parcelle. La modélisation de l'évolution de la parcelle agricole est réalisée à l'aide du formalisme des automates temporisés. La démarche consiste à confronter une suite d'observations, issues des images, avec une suite d'états, proposés par la simulation du système dynamique, dans le but de restreindre le nombre d'états susceptibles de représenter l'occupation du sol. Les automates temporisés sont généralement employés pour la représentation des systèmes temps-réel mais s'adaptent bien dans ce cadre puisqu'ils permettent l'expression des contraintes temporelles et des cycles caractéristiques de l'évolution de la parcelle agricole. Les occupations du sol correspondent aux états de l'automate reliés par des transitions munies de contraintes temporelles exprimées à l'aide d'horloges. La discrimination des occupations du sol consiste à comparer les observations, dérivées des images par la préclassification, à l'état attendu par la simulation de l'automate. Ce problème est abordé comme un problème de vérification et est exprimé en logique TCTL (Timed Computational Tree Logic) par une propriété d'atteignabilité. Cette vérification est ainsi réalisée, par parcelle et sur chaque couple d'images, dans une étape de prévision puis de postdiction. Le but de la postdiction consiste à raffiner les états proposés pour chaque parcelle dans les images précédentes compte-tenu des derniers résultats.

La méthode a été expérimentée sur une série de cinq images (aériennes, Landsat TM et Spot). La description de l'automate puis la vérification des propriétés d'atteignabilité a été réalisée à l'aide de Kronos [Yov97]. Les résultats sont encourageants, le taux de reconnaissance a augmenté sur la plupart des images : les taux variant de 63 à 95% après la préclassification atteignent 70 à 95% après les traitements, et le nombre de parcelles ambiguës a fortement diminué (jusqu'à 52,6% pour une image).

Actuellement nous étudions l'introduction de probabilités dans la modélisation par automates temporisés. Nous pourrons alors caractériser quantitativement les solutions issues de la simulation et fusionner ces résultats avec ceux de la préclassification. Un critère de décision pourra ainsi être appliqué si une ambiguïté subsiste à la suite du raisonnement. Nous envisageons également de valider notre méthode sur des images de la région Lorraine.

6.2 Apprentissage automatique et structuration de données

Participants : Catherine Belleannée, Laurent Blin, François Coste, Daniel Fredouille, Israël-César Lerman, Yoann Mescam, Konan Lemée [bourse MENRT], Laurent Miclet, Jacques Nicolas, Valérie Rouat, Basavaneppa Tallur, Raoul Vorc'h.

[Yov97] S. YOVINE, « Kronos: A verification tool for real-time systems », *International Journal of Software Tools for Technology Transfer 1*, 1997.

Mots clés : inférence grammaticale, analyse de données, classification automatique.

Résumé : *L'automatisation de la construction de modèles de systèmes complexes est au cœur des motivations des recherches effectuées ici. Nous focalisons nos travaux pour le traitement de données qui se présentent sous forme de séquences discrètes finies. L'analyse de ces séquences passe généralement par deux étapes : une étape de prétraitement d'analyse à un niveau lexical et éventuellement syntaxique, où il faut regrouper les séquences ou sous-séquences similaires, et une étape d'inférence grammaticale qui conduit au modèle souhaité.*
Nous traitons également des problèmes importants associés au développement pratique de ces outils : la réduction de la complexité d'un système descriptif et la comparaison de modèles structurant un même ensemble d'objets.

6.2.1 Analyse de séquences

Participants : Catherine Belleannée, Daniel Fredouille, Israël-César Lerman, Yoann Mescam, Jacques Nicolas, Basavaneppa Tallur.

Classification de protéines La classification de protéines repose de façon déterminante sur l'emploi de matrices de similarités entre acides aminés qui traduisent les ressemblances élémentaires entre les lettres des séquences. Afin d'améliorer les résultats des classifications, nous avons travaillé dans deux directions : l'automatisation du choix de la matrice la plus pertinente et le processus même de construction de ces matrices.

En ce qui concerne le premier point, nous avons continué à travailler sur la classification de protéines de la famille MIP, en collaboration avec les biologistes de l'Upresa 6026 à Rennes, avec pour objectif d'essayer de prédire leur fonction par la classification des segments de séquences pertinents. Nous avons montré que notre méthode de classification hiérarchique permettait de bien séparer les deux fonctions, à savoir : AQP (aquaporines spécialisées en transport de l'eau) et GLPF (transport de petits solutés tels que le glycérol) à condition d'avoir «bien» choisi la matrice des similarités entre acides aminés. Pour cela, nous utilisons le critère de la *statistique globale* et la connaissance d'une classification partielle pour sélectionner de manière automatique la meilleure matrice pour la prédiction [30, 29]. Nous cherchons actuellement à réduire l'influence d'un autre facteur important pour la qualité de la classification, qui correspond à la sélection de zones particulières homologues dans l'ensemble des séquences. Nous explorons deux voies complémentaires pour déterminer ces zones. La première consiste à réutiliser le critère de la statistique globale pour ce nouveau réglage. La seconde consiste à utiliser un modèle syntaxique de l'organisation des séquences (mis au point à partir des connaissances biologiques) pour trouver des alignements suffisants pour la classification.

Le deuxième axe des recherches en classification de protéines s'est intéressé à la construction des matrices de similarité. Dans le cas de séquences déjà alignées, notamment à partir de considérations structurelles, nous[LPR94] avions analysé le mécanisme de calcul des matrices

[LPR94] I.-C. Lerman, P. Peter, J. Risler, « Matrices AVL pour la classification et l'alignement de séquences protéïques », *Rapport de Recherche n° 2466*, Inria, 1994.

d'association entre acides aminés de Dayhoff et des Henikoffs (matrices «Blosum»). Nous en avions déduit des matrices conformes à l'optique AVL, ainsi que l'adaptation de la méthode AVL à la classification hiérarchique d'un ensemble de séquences protéiques d'une famille donnée. Des expériences comparatives concluantes ont été menées relativement à des familles telles que les Cytochromes, les Globines et les Synthétases. Cependant, nous dépendions alors de matrices publiées dans la littérature et établies à partir de l'alignement d'un très grand ensemble de séquences qui pouvaient ou non comprendre les séquences traitées.

L'idée du nouveau travail que nous venons de mener consiste à reprendre la construction en reconsidérant certains de ses paramètres et surtout, en induisant la matrice de similarité entre acides aminés de façon intrinsèque, directement à partir de la famille de séquences alignées, qu'il s'agit d'organiser en classes et sous classes de proximité. Les résultats obtenus correspondent à un progrès par rapport à ceux de 1994. Le travail mené s'est inscrit dans le cadre d'un stage de DESS en Mathématiques appliquées de l'université de Rennes 1, de Sébastien Josse (septembre 1999).

Enfin, nous avons une collaboration avec l'université de Postdam, qui s'est concrétisée par une invitation dans notre projet de deux personnes pour une semaine, sur la classification de données d'expressions d'ARN pour différents mutants de gènes de plantes. Nous étudions l'applicabilité de techniques classiques d'analyse de correspondances sur ce problème.

Recherche de motifs dans les séquences Nous avons terminé un travail de recherche du motif d'initiation de la traduction chez *E. coli*, qui a donné lieu à publication dans un journal [19]. L'approche a consisté à coupler une recherche combinatoire de motifs de type disjonction de présence/absence de mots dans des intervalles de positions fixés avec une construction d'arbre de décision dont chacune des variables représente le nombre d'occurrences des motifs préalablement trouvés. Sur le génome complet, des taux de reconnaissance approchant les 90% ont pu être observés, ce qui en fait une méthode performante qu'il faudrait maintenant valider sur d'autres génomes.

Nous participons à une action de recherche concertée Inria, Remag, portant sur la recherche de motifs dans les séquences, en collaboration avec l'institut Pasteur à Paris, l'université de Genève, le projet Polka à Nancy, l'université de Marne La Vallée et le projet Algo à Rocquencourt. Nous avons donné deux séminaires dans ce cadre et un premier travail de prédiction de promoteurs chez les prokaryotes a été mené à terme, qui a donné lieu à la présentation d'un poster à la conférence «Data Mining for Bioinformatics». Nous utilisons une approche originale par inférence grammaticale (voir paragraphe suivant), qui nous a permis de réduire un automate initial de plusieurs millions d'états à une centaine. Les résultats en prédiction semblent pour l'instant comparables aux autres méthodes, mais il existe de nombreuses pistes d'améliorations, dont une, l'inférence d'automates non déterministes, est le point de départ d'une nouvelle thèse (D. Fredouille).

Signalons que nous avons également démarré une collaboration avec le laboratoire de l'Inserm Germ de Rennes (unité U435) sur la conception d'un assistant logiciel pour la constitution et la mise à jour automatique d'une base de données de séquences. Bien que ce travail ne ressorte pas directement de la recherche de motifs, il constitue un préalable indispensable à celle-ci. Nous avons commencé en parallèle un travail de développement sur l'outil Forest dont un premier prototype avait été construit lors de la thèse de R. Gras. Cet outil exploratoire de

motifs, basé sur la construction d'arbres de suffixes sur des génomes complets, sera testé sur les données du Germ et mis à disposition dans le cadre de l'action Remag. Il fait l'objet d'une première collaboration avec l'université de Genève.

6.2.2 Inférence grammaticale

Participants : Laurent Blin, François Coste, Daniel Fredouille, Konan Lemée [bourse MENRT], Laurent Miclet, Jacques Nicolas.

Nous avons proposé et formalisé l'extension du cadre de l'inférence grammaticale à l'inférence d'automates classifieurs. Ce travail s'est poursuivi dans le cadre de la thèse de F. Coste, qui sera soutenue en janvier 2000. Considérer l'inférence d'automates classifieurs permet d'envisager un champ d'application plus vaste, mais aussi une meilleure compréhension et formalisation de l'espace de recherche pour l'inférence d'automates à partir d'exemples positifs et négatifs. Dans cette nouvelle approche, nous pouvons considérer l'inférence de plusieurs langages simultanément, tout en contrôlant leur intersection. L'inférence d'automates à partir d'exemples positifs et négatifs peut ainsi. par exemple, être considérée comme l'inférence d'un langage positif et d'un langage négatif sous la contrainte d'une intersection vide. Dans ce cadre, il est alors possible de généraliser les algorithmes par fusion classiques et de caractériser l'ensemble des fusions incompatibles par une modification du critère d'équivalence de Nerode. Introduisant un nouvel espace de recherche implicite portant uniquement sur les paires d'états, nous proposons une généralisation des algorithmes par fusion sous la forme de deux algorithmes de Branch and Bound.

Cette approche a été implémentée au sein de la plate-forme logicielle générique pour l'inférence de machines à états finis commencée l'année précédente et dont le développement a été poursuivi. Le but à moyen terme est de pouvoir proposer le logiciel de manière publique. Les premiers tests indiquent les bonnes performances de l'approche proposée sur des exemples artificiels. Une comparaison systématique des divers facteurs d'efficacité des algorithmes proposés est en cours et devrait être poursuivie sur des données réelles.

Nous avons également développé un travail sur le traitement de l'inférence grammaticale par unification décrit dans [33]. Ce travail comprend deux volets :

- Un volet porte sur les algorithmes classiques d'inférence d'automates par fusion d'états. Nous avons adopté une représentation par arbres infinis des automates. L'implémentation d'un nouvel algorithme, Aimes, tirant parti des mécanismes de retour-arrière de Prolog a permis d'atteindre les résultats des meilleurs algorithmes actuels d'inférence d'automates sans utilisation d'une composante stochastique. Nous avons pu traiter un problème de grande taille sur les séquences génétiques de cette façon (voir recherche de motifs dans les séquences).
- Un second volet résulte d'un travail avec Christian Retoré et P. Sébillot sur l'analyse du langage naturel. Christian Retoré (cf. rapport du projet Paragraphe) collabore avec le projet Aïda depuis son arrivée à Rennes en octobre 97, une partie importante de ses recherches concernant les grammaires catégorielles et le traitement des langues naturelles. Nous avons étudié les algorithmes d'apprentissage des grammaires catégorielles (AB ou «basiques», principalement rigides et k-valuées) et encadrons la « Tesi di Laurea » de Ro-

berto Bonato (Università di Verona, échange Erasmus) sur ce thème. Le but est d'étendre l'apprentissage dans le cadre du calcul plus puissant de Lambek.

Enfin, nous avons mené une étude sur l'inférence d'automates déterministes par fission, c'est à dire en procédant par spécialisation d'un automate initial universel. Ce travail a fait l'objet d'un stage de DEA et mis en valeur l'absence de dualité avec le problème de la fusion, notamment de par la nécessité de considérer des automates non déterministes.

6.2.3 Apprentissage de structures d'arbres

Participants : Laurent Blin, Laurent Miclet.

Le domaine des données structurées en arbres est relativement peu abordé par les recherches en apprentissage. Lorsqu'un arbre est choisi pour représenter des données que l'on veut manipuler, c'est habituellement pour profiter de propriétés hiérarchiques qu'il apporte implicitement entre ses diverses composantes. Le problème vient de ce que la plupart des techniques d'apprentissage ne permettent pas leur manipulation directe. Il faut auparavant «mettre à plat» ces données, ce qui entraîne la perte des informations qu'induisait l'arborescence. Maintenir ces structures intactes au cours d'un apprentissage permettrait la conservation et l'utilisation de leurs propriétés. Pour cela, l'adaptation ou la création de nouvelles techniques d'apprentissage est nécessaire.

Les travaux actuellement en cours portent sur des apprentissages simples, de la famille des K-plus proches voisins, et sur les diverses notions de distances entre arbres. L'application envisagée pour ces travaux est l'apprentissage automatique de la prosodie pour la synthèse vocale de phrases du français. Les données y sont des phrases, représentées principalement sous la forme de leur arbre syntaxique et d'informations pragmatiques, aux feuilles duquel sont positionnés les différents phonèmes avec leurs caractéristiques prosodiques. Un ensemble de phrases étiquetées a été créé à cet effet en collaboration (non contractualisée) avec le Cnet de Lannion. Ce corpus (en langue française) permettra de tester divers passages du numérique vers le symbolique, ainsi que les résultats des méthodes d'apprentissage mises en œuvre.

Un séjour scientifique de 6 mois de L. Blin au laboratoire Star[4] de SRI International[5] (Cambridge, Angleterre, et Menlo Park, Californie, États Unis) se déroule actuellement pour implémenter et tester une nouvelle approche basée sur le concept d'analogie. La prédiction des caractéristiques prosodiques d'une phrase X_1 s'effectue par la recherche dans un ensemble de phrases (de caractéristiques prosodiques connues) de la phrase Y_1 la plus proche (au sens d'une distance définie sur les caractéristiques arborescentes choisies). L'hypothèse de base de cette approche est qu'il doit être possible de trouver dans cet ensemble de phrases un couple d'individus (X_2,Y_2) présentant la même alternance structurelle que le couple (X_1,Y_1), et d'en déduire la modification des caractéristiques prosodiques induite par cette modification des caractéristiques structurelles. Les premiers tests de cette approche seront effectués sur un ensemble de phrases issues du Boston University Speech Radio Corpus, de langue anglaise.

4. Speech Technology And Research laboratory.
5. Stanford Research Institute International.

6.2.4 Réduction de la complexité d'un système descriptif

Participants : Israël-César Lerman, Valérie Rouat, Raoul Vorc'h.

Analyse combinatoire des données dans le problème de la satisfiabilité Le problème SAT est celui de la satisfiabilité d'un ensemble de clauses. Le problème de la détermination du nombre de solutions d'une instance SAT, noté #SAT est un problème #P-complet. Il est bien connu qu'il n'existe pas, jusqu'à présent, d'algorithme déterministe polynômial capable de résoudre ces problèmes. Même plus, il n'existe pas un tel algorithme pouvant assurer une approximation du nombre de solutions dont la précision est supérieure à un seuil donné. Notre approche est différente et correspond à une optique de recherche opérationnelle où il s'agit de trouver la meilleure approximation, quitte ensuite à évaluer sa qualité. À cette fin, nous utilisons une algorithmique issue de l'analyse combinatoire des données, fondée sur une représentation ensembliste, géométrique et logique d'une instance SAT. On se place pour cela dans l'espace des interprétations et on établit une bijection entre une clause et l'ensemble des interprétations qui la falsifient. Il s'agit d'un «cylindre ponctuel» dans l'espace géométrico-logique $\{0,1\}^N$, où N est le nombre de variables. Une telle représentation permet d'une part, une vision synthétique des algorithmes proposés dans la littérature et, d'autre part, la prise en compte de caractéristiques statistiques globales de l'instance. Il en résulte la spécification d'algorithmes performants pour le calcul approché du nombre de solutions dans notre cas.

L'utilisation de la classification et de l'analyse combinatoire des données permet de procéder conformément au principe général *diviser pour résoudre*. Nos résultats actuels portent sur le problème 3-SAT où chaque clause comprend trois variables. On se place au pic de la difficulté (rapport nombre de clauses/nombre de variables égal à 1,2) dans le cas d'un modèle aléatoire uniforme, sans aucune structure stochastique cachée (les trois variables sont choisies uniformément au hasard). Le faible nombre de variables par clause nous conduit à adopter une technique de sériation. Relativement au tableau réorganisé Clauses×Variables, une formule d'indépendance approchée ramène le problème à la meilleure coupure en deux du tableau, par déplacement linéaire. Des critères de coupure de complexité polynômiale (d'ordre 1 ou 2) en le nombre de clauses ont été étudiés. Le plus élaboré tient intimement compte de la structure statistique de l'instance aléatoire réorganisée. Les résultats expérimentaux déjà obtenus sont tout à fait prometteurs. Ils améliorent sensiblement les résultats précédemment obtenus dans la littérature. La recherche doit se poursuivre encore pour un temps, afin d'améliorer encore la qualité de l'approximation [20].

Ces travaux ont conduit à la soutenance de la thèse de Valérie Rouat [15].

Classification hiérarchique de gros ensembles sous contrainte de contiguïté, application à l'imagerie médicale Il est maintenant bien admis et depuis longtemps que la technique de recherche des plus proches voisins réciproques est cruciale pour la conception d'algorithmes de construction ascendante hiérarchique d'arbres de classification sur de «gros» ensembles. La situation spécifique considérée et qui se retrouve dans nombre d'applications est celle où, pour la formation des classes, une contrainte de contiguïté doit être respectée. On suppose de plus que le nombre d'objets contigus à un objet donné reste limité par une

constante fixée à l'avance. C'est typiquement la situation pour la classification des pixels d'une image numérisée. K. Bachar [ESSCA, Angers] avait, notamment dans le cadre de sa thèse [université de Rennes 1, décembre 1994], élaboré et analysé sur les plans théorique et expérimental, un algorithme CAHCVR (Classification Ascendante Hiérarchique sous Contraintes utilisant les Voisins Réciproques). On démontre et on vérifie dans la pratique que la complexité moyenne en temps de calcul devient linéaire, au lieu de quadratique dans le cas général, en fonction du nombre d'objets, ce qui est optimal.

Dans ces conditions, il importait d'approfondir l'étude algorithmique par rapport aux critères de type inertiel mis en œuvre. Il s'agissait aussi de valider la méthode en l'expérimentant sur des images difficiles et sensibles fournies par la radiologie. Il convient enfin d'étudier l'influence théorique et pratique de cette algorithmique sur des critères de dissimilarité «informationnelle» issus de la méthode AVL de la vraisemblance des liens. À cet égard, les programmes sont en cours de validation sur des données réelles.

C'est ainsi que s'est instaurée une collaboration avec le laboratoire LTSI de l'université de Rennes 1, dirigé par Jean Louis Coatrieux. Les résultats déjà obtenus sont fortement encourageants.

Comparaison d'arbres de classification On peut distinguer deux formes de comparaison : la première, globale, au moyen d'un coefficient d'association entre les deux arbres pris dans leur ensemble et la seconde, locale, où il s'agit de déterminer des associations entre une partition P observée sur le premier arbre et une partition Q observée sur le second.

Un travail important, combinatoire, statistique et informatique a déjà été mené relativement à la première forme de comparaison. Rappelons que par rapport à notre méthode de construction d'un coefficient d'association entre deux structures combinatoires sur le même ensemble d'objets, deux approches sont proposées et étudiées. Les deux approches se basent sur la représentation d'un arbre de classification au moyen d'une préordonnance ultramétrique. Pour la première, le codage se fait en utilisant au niveau de l'ensemble P2 des paires d'objets, la notion de rang moyen. Pour la seconde, la préordonnance est représentée par son graphe au niveau de P2×P2. La première partie de ce travail où les différents points de vue présentés dans la littérature sont analysés pour bien situer notre apport a conduit à [21].

Il importe également, comme nous l'avons mentionné ci-dessus, de pouvoir comparer deux arbres, mais de façon «plus locale», partition à partition. Le problème réel s'est présenté dans le cadre de la classification de 2684 parcelles agricoles, formant un bassin versant qui alimente en eau la ville de Rennes. La zone est photographiée en 256 niveaux de gris. Des raisons de compression initiale de l'information, mais aussi de stabilité des résultats, nous conduisent à réduire l'échelle de variation de la luminance et à comparer deux arbres de classification AR1 et AR2, respectivement issus d'une échelle à 128 niveaux de gris et d'une échelle à 64 niveaux de gris. Relativement au problème méthodologique général, des techniques spécifiques ont été développées concernant :

- la collecte d'un couple de partitions (P,Q), consistantes dans leur comparaison en termes de tailles des classes, où les classes de P (resp. Q) correspondent à des noeuds de AR1 (resp. AR2) ;

– la mise en correspondance des deux partitions P et Q et l'utilisation de différents coefficients de comparaison entre P et Q, dont principalement celui - se référant à notre approche et conçu dans un contexte relationnel - dit CCR (Coefficient Centré Réduit).

Le comportement de ce coefficient pour des partitions qui se ressemblent beaucoup a été très intéressant à constater. On se trouve dans des situations de nette discrimination dans l'échelle de la «grande ressemblance». Ces travaux sont appelés à être poursuivis, notamment pour mieux comprendre le comportement du coefficient CCR relativement à la vérité terrain, en ce qui concerne l'occupation du sol.

Ces travaux ont donné lieu au stage de DESS «génie mathématique en calculs scientifique et statistique» (université de Franche-Comté) de D. Morel (juin 1999). Ils ont été effectués dans le laboratoire d'informatique de l'Ensar, en collaboration avec G. Douaire, directeur du laboratoire.

6.3 Ingénierie de la langue

Participants : Israël-César Lerman, Jacques Nicolas, Ronan Pichon, Pascale Sébillot.

Résumé : *Dans le cadre du développement de méthodes et outils linguistiques permettant d'augmenter les possibilités d'apparier une requête et les textes d'une base documentaire, nous nous sommes intéressés cette année à l'acquisition automatique de lexiques sémantiques à partir de corpus, en étudiant deux points : l'acquisition d'éléments du lexique génératif de Pustejovsky par des méthodes de programmation logique inductive et l'acquisition de lexiques basés sur la sémantique componentielle de Rastier par des méthodes de classification.*

Une autre voie de recherche théorique est poursuivie, en collaboration avec C. Rétoré du projet Paragraphe, sur les liens existant entre le formalisme des grammaires catégorielles et le programme minimaliste de N. Chomsky. Une correspondance peut être établie entre l'analyse par une grammaire d'arbres et une démonstration dans le calcul de Lambek. Le problème du déplacement de constituants (opération move *de Chomsky) est résolu par un étiquetage de la démonstration comportant à la fois des informations phonologiques et sémantiques.*

Notre point de départ est le système que nous avons réalisé (cf. [24]), qui permet de déterminer automatiquement la relation qu'entretiennent les constituants d'une séquence binominale de la forme *Nom Nom* en anglais et *Nom à/de (déterminant) Nom* en français, en se basant uniquement sur la forme de la séquence et sur la sémantique des mots qui la composent. Pour les composés contenant un constituant déverbal (*truck-driver, séquençage de l'*ADN), notre calcul automatique se base sur la satisfaction de la structure argumentale du prédicat verbal sous-jacent. Les composés sans constituant déverbal sont traités en généralisant la notion d'attachement d'information prédicative aux noms simples, en faisant appel à une représentation lexicale élaborée intégrant des informations pragmatiques telle que la met en œuvre Pustejovsky dans le lexique génératif. Dans ce formalisme, la structure des *qualia* représente un

mot en termes de rôles sémantiques – fonctionnel, agentif, constitutif, formel[6] – qui rendent explicites les différents éléments de sens nécessaires à sa définition. Pour un nom, ces rôles sont fréquemment tenus par des verbes.

Comprendre ces séquences complexes constituant les éléments porteurs de l'information clé dans un domaine permet d'augmenter les possibilités de détecter des concepts équivalents entre une requête et une base de documents indexés. Il s'agit, non seulement, de rechercher des formulations voisines de celles utilisées en réalisant l'expansion des index représentant les textes, par ajout de synonymes, de variantes morpho-syntaxiques, etc., mais aussi, et c'est le point qui nous intéresse ici, en générant des variantes en privilégiant plusieurs types de liens sémantiques, ce pour quoi le modèle du lexique génératif s'avère particulièrement adapté (cf. [24] pour une expérience sur ce point). Ainsi, si une requête contient la séquence *jaugeur de carburant*, le fait de disposer d'un lien entre le nom *jaugeur* et le verbe *mesurer* (fonction typiquement associée) permet, par exemple d'étendre la recherche aux séquences voisines *mesure du carburant* ou *mesurer le carburant*.

Cependant, et ce quel que soit le cadre théorique choisi, on ne dispose pas de lexiques sémantiques suffisamment précis quel que soit le domaine de l'application étudiée. Nous nous intéressons donc à l'acquisition automatique, à partir de corpus de textes, de deux types de lexiques sémantiques : des lexiques respectant le formalisme de Pustejovsky, dans le cadre d'une action de recherche partagée de l'AUPELF-UREF de deux ans débutée cette année, en collaboration avec Pierrette Bouillon (Issco Genève), Cécile Fabre (ERSS Toulouse) et Laurence Jacqmin (université de Bruxelles) (cf. 7.7), et des lexiques suivant la sémantique componentielle de Rastier. Pour le premier type de lexiques, nous nous sommes intéressés, dans un premier temps, à mettre au point une méthode d'apprentissage automatique de type programmation logique inductive (Pli) permettant, dans un corpus de textes techniques étiquetés catégoriellement, de distinguer automatiquement les couples nom-verbe (N-V par la suite) liés par une relation sémantique codée dans le lexique génératif des autres couples N-V. Plus précisément, nous avons utilisé les outils d'étiquetage statistiques développés par l'Issco (outils Multext) sur un manuel de maintenance d'hélicoptères (moins de 2% d'erreurs d'étiquetage catégoriel obtenus). Les données d'apprentissage ont été obtenues de la façon suivante : pour tous les noms du corpus, les dix verbes qui leur sont le plus associés selon la méthode du Khi2 sont retenus et les paires obtenues sont annotées manuellement comme pertinentes ou non pertinentes par rapport à ce que l'on veut apprendre (c'est-à-dire la structure des qualia). Les exemples positifs et négatifs sont constitués à l'aide des contextes d'apparition de ces N et V dans les phrases (la catégorie grammaticale du mot avant et après le nom, celle du mot avant le verbe, le type de verbe, la distance en nombre de verbes conjugués et la position du verbe par rapport au nom). Les 4000 exemples positifs et 7000 exemples négatifs automatiquement générés sont fournis en entrée de Progol, mise en œuvre de la Pli développée par Muggleton, qui produit alors des clauses par généralisation de certains exemples positifs. Dans notre cas, les clauses générales obtenues couvrent 88% d'exemples positifs et seulement 5% d'exemples négatifs (coefficient de Pearson de 0.84). Nous avons également validé empiriquement la méthode d'apprentissage en utilisant ces clauses générales pour étiqueter les couples N-V du corpus de départ et en comparant la

6. Le rôle fonctionnel indique la fonction typique de l'objet dénoté, l'agentif le mode de création, le constitutif ses éléments constitutifs et le formel sa catégorie sémantique.

pertinence des décisions prises par rapport à un étiquetage manuel et par rapport à la liste des couples fortement corrélés proposés par le test du Khi2. La proportion, pour un certain nombre de noms significatifs du corpus, de verbes retrouvés qui participent effectivement à leur structure qualia est meilleure que celle obtenue par le Khi2 (83% contre 64% dans notre test). Les résultats sont cependant encore bruités et nous travaillons à leur amélioration par un étiquetage sémantique du corpus.

Parallèlement à ce travail, nous nous intéressons à l'acquisition automatique de lexiques sémantiques basés sur la sémantique componentielle (SC) de Rastier, théorie linguistique dans laquelle l'accent est mis sur les relations entre les significations des mots au sein d'un lexique, et dont une thèse est que ces relations sont fortement dépendantes d'observations d'utilisation des mots en corpus. Dans SC, la signification d'un mot est définie par les différences qu'elle entretient avec les autres significations présentes dans le lexique. Ces différences sont représentées par des sèmes (ou traits sémantiques). Au sein d'une même classe sémantique, correspondant à un groupe de mots partageant certains traits sémantiques et pouvant être échangés dans certains contextes, les éléments possèdent des sèmes génériques correspondant aux contextes dans lesquels ils peuvent effectivement être échangés, et des sèmes spécifiques, correspondant aux autres contextes. Pour Rastier, le sens d'un mot est totalement déterminé par le contexte qui l'entoure, et deux types de contextes sont fondamentaux pour caractériser les relations de signification lexicales : le thème de l'unité de texte dans laquelle est située l'occurrence étudiée et son voisinage.

La présence d'un thème peut être caractérisée par la co-présence, dans une unité de texte, de quelques mots typiques de ce sujet. Nous avons donc mené une expérience visant à déterminer automatiquement les sujets abordés dans un corpus du Monde Diplomatique (corpus global de 7.8 millions de mots, dont 1 million ont servi à cette expérience) étiqueté catégoriellement, et à calculer les listes des mots caractéristiques de ces thèmes. Pour ce faire, nous avons sélectionné les noms apparaissant assez fréquemment (au moins huit fois) dans le sous-corpus, et avons adjoint à ces 165 noms présents dans 8570 paragraphes un vecteur formé de leur lemme et des numéros des paragraphes dans lesquels ils apparaissent. En utilisant une méthode de classification hiérarchique, l'analyse de la vraisemblance du lien (AVL), avec pour mesure de proximité la distribution relative des noms à travers les paragraphes, nous avons mis en évidence un ensemble de thèmes traités dans le corpus, tels que : presse, institutions, finance, situations de crises, etc., chacun étant associé à un petit ensemble de mots-clés déterminés automatiquement (par exemple le thème presse est représenté par l'ensemble *{journaliste, journal, presse}*). Les résultats ont été évalués par cinq personnes qui, parmi les 80 ensembles de mots proposés (45 en prenant en compte le facteur discriminant de l'AVL) en ont validé (accord des cinq personnes) 27 (resp. 21). Ces ensembles validés ont été comparés à l'index accompagnant le CD-Rom du corpus : 92% des thèmes validés sont présents dans le corpus, et ces thèmes recouvrent 60% des paragraphes du corpus total [28].

Nous nous sommes ensuite intéressés à l'étude de la variation des interprétations des différentes occurrences des mots selon les thèmes des unités de textes dans lesquelles elles apparaissent. Nous avons, pour chaque thème reconnu, construit un sous-corpus des paragraphes lui correspondant. Pour un nom donné, nous avons construit un vecteur de voisinage formé des noms et adjectifs apparaissant dans une fenêtre de 5 mots avant et après chacune de ses occurrences dans les paragraphes du thème étudié. Nous avons ensuite étudié les similarités

et dissimilarités de sens entre deux occurrences d'un même mot dans deux thèmes différents, ou entre deux mots dans un même thème, en détaillant les similarités et dissimilarités entre leurs vecteurs de voisinage. Cette étude se fait par calcul de l'intersection ou de la différence ensembliste entre ces vecteurs de voisinage, et nous avons interprété les ensembles de mots ainsi obtenus en y recherchant des séquences caractérisant une différence entre la signification de mots. Les résultats obtenus ont permis de bâtir une représentation partielle des significations des mots étudiés dans les thèmes abordés (sous forme de tableau ou de réseau sémantique dans lequel un arc étiqueté /sème/ entre deux nœuds A et B signifie que le sème /sème/ participe à la signification de A et pas à celle de B) [27]. La mise en évidence de ce qui rassemble et distingue des occurrences d'un même mot dans deux thèmes, basée sur le calcul des intersections et différences ensemblistes, est entièrement automatisée. Seule l'interprétation, c'est-à-dire le nommage de la ressemblance ou de la différence mise au jour, est manuelle. Pour ce qui concerne la mise en évidence de points communs et d'éléments distinctifs entre des mots différents dans un même thème, seules les sous-listes caractérisant une différence particulière doivent encore être extraites (et interprétées) manuellement des contextes associés aux mots. Nous travaillons actuellement à l'automatisation de cette tâche et à la structuration des vecteurs de contexte.

6.3.1 Grammaires minimalistes et logique linéaire

Edward Stabler a proposé en 96 une formalisation du programme minimaliste de Chomsky sous la forme de grammaires d'arbres. Ce sont des grammaires totalement lexicalisées, où chaque entrée est une liste de traits, et où l'assemblage des constituants (*merge*, opération binaire) ainsi que leur déplacement à l'intérieur d'un arbre déjà constitué (*move*, opération unaire) sont gérés par l'annulation de traits de polarité opposée. Cela a permis à C. Retoré (cf. le rapport du projet paragraphe pour les détails et les publications), en collaboration avec A. Lecomte (université de Grenoble II et projet Calligramme) de préciser et de formaliser le lien entre grammaires catégorielles et programme minimaliste, en faisant un autre usage grammatical du calcul de Lambek que celui proposé initialement. Les entrées lexicales *à la Stabler* sont traduites par des formules du calcul de Lambek, et l'on doit obtenir une démonstration dans le calcul de Lambek du symbole initial de la grammaire. Bien sûr, la difficulté est de rendre compte des déplacements de constituants ;cela est fait en étiquetant la démonstration. Les étiquettes sont des chaînes qui comportent à la fois les informations phonologiques (les mots et les traits d'accord) et les informations sémantiques (à la Montague : variables, quantificateurs, ...). Ce genre de travaux intéresse les deux domaines rapprochés : d'une part les grammaires catégorielles peuvent ainsi profiter de la profondeur d'analyse de la théorie chomskienne, et d'autre part les grammaires minimalistes acquièrent ainsi une interface plus aisée avec la sémantique, par exemple la sémantique de Montague.

La connexion entre calcul de Lambek et grammaires minimalistes n'est qu'un aspect du rapprochement de l'approche générative (grammaires transformationnelles) et de l'approche logique (grammaires catégorielles). C'est pourquoi C. Retoré et Edward Stabler (UCLA, Los Angeles) ont organisé un workshop sur « Resource logics and minimalist grammars » dans le cadre de l'école d'été « European Summer School in Logic, Language and Information » et ont écrit à ce propos un article de synthèse.

6.4 EIAO (Assistants intelligents pour l'enseignement)

Participants : Jacques Nicolas, Dominique Py, François-Gilles Carpentier, Romuald Texier.

6.4.1 Individualisation des logiciels de formation

Nous avons débuté cette année une collaboration avec la société IDP (cf section 7.5) qui développe des dispositifs de formation professionnelle pour adultes. La thèse de Romuald Texier porte sur l'intégration d'assistants intelligents au sein des logiciels d'auto-formation. L'objectif est de proposer une approche générique apprentissage/coopération permettant la conception d'outils génériques d'aide à l'apprenant, pour les logiciels de formation. La proposition consiste à construire ces outils sous la forme d'un ou plusieurs assistants, possédant des capacités d'apprentissage et de coopération, qui observeront les actions de l'apprenant afin d'élaborer un modèle de son comportement et exploiteront ce modèle pour lui apporter une assistance adaptée, ou bien pour produire une synthèse des sessions destinée au formateur.

6.4.2 Interaction dans les EIAO de calcul formel

L'arrivée d'outils de calcul formel fiables et performants dans l'enseignement des mathématiques pose la question de leur adéquation aux situations pédagogiques. Dans le cadre d'un projet piloté par l'INRP (cf section 7.4), nous nous intéressons à la conception d'environnements d'apprentissage basés sur des outils de calcul formel, et à la modélisation de l'interaction au sein de ces environnements.

Le groupe de Rennes travaille plus précisément sur le domaine de l'étude des variations d'une fonction réelle. Dans un premier temps, nous avons analysé l'interaction d'élèves de première avec le logiciel Derive dans le cadre de tâches d'étude de fonctions, et nous avons montré comment l'élève navigue entre différents registres (graphique, numérique, symbolique, papier-crayon).

Cette année, notre groupe a poursuivi ses travaux par la spécification et la réalisation d'un environnement logiciel dédié à l'étude des variations. L'architecture générale a d'abord été étudiée. Deux possibilités ont été envisagées : dans la première, l'outil de calcul formel est «caché» par le logiciel, dans la seconde, l'élève manipule directement l'outil de calcul formel tout en bénéficiant d'une aide et d'un guidage logiciels. Nous avons retenu la première architecture et choisi de réaliser une maquette limitée à l'étude des variations, sur un intervalle fini, d'une fonction rationnelle définie et continue sur cet intervalle, à l'aide de la dérivation. Dans cette maquette, l'élève a pour tâche de remplir un tableau de variations interactif, en s'aidant d'outils d'exploration et de transformation, puis de justifier les valeurs entrées dans ce tableau.

Une première version (sans le module de justification) a été expérimentée par entretien individuel auprès de huit élèves. Les résultats de l'expérimentation sont en cours d'exploitation. En première analyse, le logiciel offre à l'élève un cadre de travail qu'il s'approprie aisément et qui lui permet de mener l'étude d'une fonction assez complexe dans un temps raisonnable, avec des gestes variés dans les différents registres.

6.5 Raisonnements et logiques non classiques

Participants : Yves Moinard, Raymond Rolland, Jean-Marc Guinnebault.

Glossaire :
circonscription logique de modèles minimaux particulière décrivant précisément l'ajout automatique d'axiomes formalisant la notion d'exception.
inférence préférentielle logique de modèles minimaux étendue où on s'autorise à considérer une relation non plus directement sur les modèles mais sur des états, ou copies de modèles.

Afin de finaliser les résultats déjà obtenus [10], nous avons développé ceux-ci dans deux directions.

D'une part, nous avons poursuivi une étude fine de la circonscription propositionnelle dans le but à la fois d'aider à une automatisation efficace et de faciliter la traduction de règles de sens commun en termes de circonscriptions. Ce second point a fait l'objet d'un rapport de DEA qui étend les précédents résultats au cas de plus de deux règles. En ce qui concerne le premier point, nous avons aussi précisé les résultats obtenus l'an dernier [25]. Nous avons surtout fourni une définition constructive d'un des plus petits ensembles Em de formules équivalent à un ensemble donné E, au sens où la circonscription de l'ensemble Em égale celle de l'ensemble E [32].

D'autre part, nous avons montré précisément comment des formalismes qui étaient réputés non exprimables en termes de circonscription le sont en fait d'une façon simple et naturelle. Il suffit d'utiliser une extension du vocabulaire afin de ramener toute inférence préférentielle finie ayant une propriété classique dans le domaine, la cumulativité, à une circonscription. Cette étude a aussi permis de démontrer qu'une variante de la circonscription introduite récemment, la circonscription par cardinalité, peut facilement s'exprimer en termes de circonscription classique, et réciproquement. Là encore, nos résultats devraient faciliter le calcul automatique, d'autant plus que nous avons aussi démontré [32] qu'un résultat technique supposé faciliter ce calcul, donné par la publication introduisant cette nouvelle sorte de circonscription, est en fait erroné.

Nous avons également débuté une étude de différentes logiques temporelles.

Enfin, nous avons achevé [13] une étude portant en particulier sur la logique conditionnelle. Il s'agit d'une logique modale avec un opérateur à deux arguments qui s'ajoute à l'opérateur d'implication classique. Cette étude a aussi fourni une caractérisation originale et prometteuse de la notion de non monotonie.

6.6 Planification et révision de croyances dans un système de dialogue

Participants : Yves Moinard, Philippe Besnard, Dominique Py.

Résumé : *Le but est d'améliorer un système déjà existant, dans un sens demandé par le Cnet (voir partie 7.6), créateur et utilisateur de ce système. Il faut permettre au système de mieux réagir en cas d'erreurs, ou d'évolution de la requête de l'utilisateur d'un logiciel interactif.*

Il s'agit d'interpréter les requêtes d'un utilisateur dans un système de dialogue coopératif homme-machine. Une logique modale complexe, combinant divers systèmes modaux, dont

certains très classiques comme KD45, est déjà utilisée comme langage de représentation des connaissances. Des méthodes d'intelligence artificielle de *révision des connaissance* et de *raisonnement par défaut* doivent donc être mises en œuvre ici. La coopération a démarré en avril 1997 pour une durée de trois ans.

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Modélisation, diagnostic et supervision de réseaux de télécommunication

Participants : Marie-Odile Cordier, Laurence Rozé, Emmanuel Mayer, Yannick Pencolé.

La convention CTI avec le Cnet concernant la surveillance de réseaux de télécommunications se poursuit en coopération avec le LIPN. La participation du projet Aïda se focalise sur deux points :

- L'acquisition automatique de scénarios de pannes discriminants. Nous avons choisi d'utiliser pour cela des techniques d'apprentissage automatique de type Pli. Le principe consiste à rechercher, pour chaque panne, un scénario discriminant qui accepte les séquences d'alarmes correspondant à la panne et rejette les séquences d'alarmes relatives aux autres pannes. Les séquences d'alarmes sont obtenues en simulant le modèle à base d'automates communicants décrivant le fonctionnement du réseau. Ce travail s'inscrit dans le cadre de la réalisation du projet Gaspar (module de discrimination).
- La construction d'automates diagnostiqueurs. Ce travail a pour objectif de construire, à partir du modèle de fonctionnement du système de gestion d'alarmes, un automate capable d'analyser les alarmes reçues par le superviseur et d'en inférer les pannes possibles. La principale difficulté est liée à la taille de cet automate et nous étudions la construction de diagnostiqueurs génériques, profitant de la structure hiérarchique du réseau, ainsi que de diagnostiqueurs décentralisés.

Ces travaux sont expérimentés sur le réseau de transmission de données Transpac ainsi que sur le réseau ATM par le LIPN. De plus, un projet a démarré dans le cadre des projets RNRT en collaboration avec Alcatel CIT, le Cnet et Ilog côté industriels, avec le LIPN/université Paris-Nord côté universitaires et au sein de l'Irisa en collaboration entre les projets Pampa, Sigma2 et Aïda.

Ce projet a pour nom Magda (Modélisation et Apprentissage pour une Gestion Distribuée des Alarmes) et a pour objectif l'étude d'une chaîne complète de supervision d'un réseau de télécommunication. Il s'agit de développer et d'expérimenter de nouvelles méthodes de gestion des alarmes et, plus précisément, de permettre une meilleure compréhension des défaillances ou des pannes, à l'aide d'outils d'acquisition d'expertise (modélisation, apprentissage), puis de reconnaître en ligne des situations à risques par des outils de corrélation d'alarmes et de diagnostic. Aïda est plus particulièrement concerné par le développement des outils de diagnostic. L'approche *diagnostiqueur* (voir 6.1.2) a été étudiée dans ce contexte et une définition de diagnostiqueurs décentralisés plus adaptés à cette application est en cours de développement.

graphes causaux temporels

7.2 Amélioration des traitements des informations temporelles dans les graphes causaux temporels

Participants : Marie-Odile Cordier, René Quiniou, Irène Grosclaude.

Le contrat EDF P21L74/B02444/0-EP881 qui avait pour objet une étude de faisabilité sur l'optimisation des traitements effectués sur les modèles de pannes causaux temporels utilisés à l'EDF, ainsi que l'application de ces modèles au problème du pronostic de l'état d'un matériel en vue d'aider à l'organisation de sa maintenance, a donné lieu à un rapport de contrat. Nous poursuivons l'étude de l'interaction des pannes dans ces graphes causaux et la possibilité de compilation de ces graphes en scénarios de pannes en nous appuyant sur les données fournies par le département Surveillance, Diagnostic et Maintenance de la DER d'EDF et relatives au diagnostic des groupes moto-pompes primaires.

7.3 Inférence grammaticale régulière pour l'apprentissage de la syntaxe en reconnaissance de la parole

Participants : Jacques Nicolas, Laurent Miclet, François Coste.

Il s'agit d'une convention CTI avec le Cnet (CCTP LAA/TSS/RCP/860) de décembre 1997 à décembre 2000. En collaboration avec l'équipe Cordial de Lannion (Irisa-Enssat) et l'université de Saint-Étienne, le projet cherche à améliorer les techniques de reconnaissance de la parole continue actuellement développées au Cnet. Il s'intéresse pour cela à un modèle de langages contraignant l'ensemble des phrases admissibles par la reconnaissance. L'objectif est de passer de la modélisation actuelle par trigramme à une modélisation par automates stochastiques réguliers construits de manière automatique par inférence grammaticale à partir d'exemples.

7.4 L'interaction dans les EIAO intégrant des instruments de calcul formel

Participante : Dominique Py.

Participation au contrat INRP «L'interaction dans les EIAO intégrant des instruments de calcul formel» dont l'objet est de concevoir des environnements d'apprentissage autour de logiciels de calcul formel.

logiciels de formation professionnelle

7.5 Développement d'assistants intelligents au sein des logiciels de formation professionnelle

Participants : Dominique Py, Romuald Texier, Jacques Nicolas.

Collaboration avec la société IDP (Ingéniérie et développement en pédagogie), à Rennes, pour le développement d'assistants intelligents au sein des logiciels de formation professionnelle. Cette coopération est matérialisée par une bourse Cifre (R. Texier).

7.6 Définition et mise en œuvre d'une théorie de la révision des croyances dans le contexte d'un dialogue coopératif

Participants : Philippe Besnard, Yves Moinard.

Contrat Cnet 97 1B 046 de mars 1997 à mars 2000 dont l'objectif est de définir une théorie de la révision des croyances pour un agent rationnel dialoguant avec un utilisateur des services audiotel. L'acquisition des modèles nécessaires passe par l'utilisation de techniques de révision des connaissances et de raisonnement par défaut.

Il s'agit de décrire précisément le processus cognitif de changement d'état mental et d'identifier les opérations logiques mises en jeu, de manière à spécifier complètement le processus logique de reconstruction des croyances.

7.7 Acquisition automatique d'éléments du Lexique Génératif pour améliorer les performances de systèmes de recherche d'information

Participante : Pascale Sébillot.

Il s'agit d'un contrat de deux ans obtenu en décembre 98 dans le cadre des Actions de Recherche Partagée de l'AUPELF-UREF, thème 1 : *Ressources Linguistiques et évaluation/outils informatiques et formalismes linguistiques*. En collaboration avec Pierrette Bouillon (Issco Genève), Laurence Jacqmin (Babel-Research Bruxelles) et Cécile Fabre (ERSS Toulouse), le projet a pour objectif de déterminer, grâce à l'utilisation de méthodes d'apprentissage sur des corpus étiquetés, les prédicats témoins des différentes facettes sémantiques des noms d'un domaine, telles qu'elles sont définies dans le Lexique Génératif de Pustejovsky, et de montrer l'apport de l'usage d'un tel lexique dans un système de recherche d'information.

7.8 Conception et contrôle de stimulateurs-défibrillateurs cardiaques intégrés

Participants : Marie-Odile Cordier, René Quiniou.

L'Action Concertée Incitative 8899 *Télémédecine et Technologies pour la Santé* du MENRT, d'une durée de deux ans, réunit le département de Cardiologie du CHU de Rennes, Ela-Recherche, le LTSI de l'université de Rennes 1 et l'Irisa. Elle a pour objectif l'amélioration des prothèses cardiaques notamment en leur apportant des capacités multisites (contrôle à partir plusieurs sondes) et multifonctions (capacité à gérer des problèmes hémodynamiques et rythmiques). Le projet Aïda est chargé d'affiner la classification des arythmies en utilisant des nouveaux électrogrammes issus d'implantation multisites de sondes et la conception de nouveaux algorithmes de contrôle basés sur la technique de reconnaissance de scénarios.

8 Actions régionales, nationales et internationales

8.1 Actions nationales

- Participation au groupe Imalaia du GdR Automatique et du GdR-PRC I3 et groupe de travail Afia (M.-O. Cordier)
- Participation à l'action concertée Remag de recherche de motifs dans les séquences génétiques (J. Nicolas, D. Fredouille) : http://www.loria.fr/projets/REMAG.
- Participation au groupe IHMC du GdR-PRC I3 (D. Py)
- Participation au groupe RàPC du GdR-PRC I3 (R. Quiniou)
- Participation au groupe de travail A3CTE : Application, Apprentissage, Acquisition de Connaissances à partir de Textes Électroniques du GdR-PRC I3 (P. Sébillot)
- Participation au groupe Colex (centre-ouest lexique) pour l'étude de la structuration d'un lexique pour l'anglais (P. Sébillot)

8.2 Réseaux et groupes de travail internationaux

- Participation au réseau d'excellence européen Monet (Model-Based and Qualitative Reasoning) (M.-O. Cordier). M.-O. Cordier est membre du « Industrial Liaison and Dissemination Committee » du réseau d'excellence européen Monet (Model-based and Qualitative Reasoning).

8.3 Relations bilatérales internationales

- Projet Procope no 99027 «Fondations pour le traitement de contradictions dans les systèmes d'information intelligents» entre l'université de Potsdam et l'Irisa (Ph. Besnard, M.-O. Cordier)

8.4 Accueils de chercheurs étrangers

- Visite de Torsten Schaub et Thomas Linke (université de Potsdam) pendant une semaine en septembre 1999.
- Invitation de Aravind Joshi (University of Pennsylvania) pendant trois jours en janvier 1999.
- Accueil de Roberto Bonato (Università di Verona, échange Erasmus), mémoire de « Tesi di Laurea » sur l'apprentissage des grammaires catégorielles (octobre 1999 – février 2000).
- Invitation de Robin Gras (Université de Genève) pendant trois jours en décembre 1999.
- Visite de Marc Dymetman et Sylvain Pogodalla (Xerox Research-Center Europe, Grenoble) deux jours en novembre 99.

9 Diffusion de résultats

9.1 Animation de la communauté scientifique

- M.-O. Cordier est co-responsable du groupe Imalaia du GdR Automatique, du GdR-PRC I3 et groupe de travail Afia.
- M.-O. Cordier participe en tant que conseillère scientifique au suivi du projet Sachem : « Système d'aide à la conduite de hauts-fourneaux » par Sollac.
- M.-O. Cordier est rédactrice en chef de RIA (*Revue d'intelligence artificielle*) et membre du comité de rédaction de AAI (*journal of Applied Artificial Intelligence*); membre du comité de programme de DX'99.
- I.-C. Lerman est éditeur associé de la revue RO-*Operations Research*, membre des comités de rédaction des revues suivantes : *Applied Stochastic Models and Data Analysis* (John Wiley & Sons) ; *Mathématique, informatique & sciences humaines* (édité par le centre d'Analyse et de Mathématiques Sociales) ; *La revue de modulad* (Editeur Inria).
- I.-C. Lerman est membre des sociétés organisatrices de l'IFCS 2000 (7th Conference of the International Federation of Classification Societies, july 11-14, 2000, Namur, Belgique.
- I.-C. Lerman est membre du comité scientifique du XII International Symposium on Applied Mathematical Methods to the Sciences, Costa Rica, january 11-14, 2000.
- L. Miclet et J. Nicolas ont été membres du comité de programme de CAP'99.
- R. Quiniou est trésorier-adjoint de l'Afia et modérateur du « bulletin électronique de l'Afia,».
- R. Quiniou est membre du comité de pilotage RFIA'2000.
- P. Sébillot a été membre du comité de programme du Workshop «Description des adjectifs pour les traitements informatiques» lors de la conférence TALN'99.
- P. Sébillot est membre du comité de lecture de la revue In Cognito.
- Invitation de I. Tellier (université de Lille) pendant trois jours en janvier 99.

9.2 Enseignement universitaire

- Option du DEA d'informatique, du DESS-ISA Ifsic et 5^e année Insa-Rennes : *représentation des connaissances* (Y. Moinard, R. Quiniou).
- Option du DEA d'informatique, du DESS-ISA Ifsic et 5^e année Insa-Rennes : *analyse des données et apprentissage* (I.C. Lerman, L. Miclet).
- Cours en Diic3 Ifsic : *images numériques : approche statistique de la reconnaissance des formes* (I.C. Lerman).
- Encadrement de projet de maîtrise Ifsic (J. Nicolas).
- Cours au DEA d'informatique de Nantes, option traitement automatique des langues : *acquisition d'informations sémantiques sur corpus* (P. Sébillot).
- Cours en $5^{ème}$ ANNÉE d'informatique de l'Insa de Rennes : *traitement automatique des langues* (P. Sébillot).

9.3 Participation à des colloques, séminaires, invitations

- Poster de J. Nicolas à la conférence «Data Mining for Bioinformatics-Towards In Silicon Biology», Hinxton (Angleterre) 10-12 Novembre 1999.
- Séminaire de J. Nicolas, École Universitaire d'Informatique de Grenoble, en Magistère sur le thème «Inférence grammaticale. Application à l'analyse de séquences biologiques», Décembre 1999. Exposé sur le même thème à l'Irisa dans le cadre du séminaire 68NQRT.

10 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] P. Besnard, M.-O. Cordier, « Explanatory Diagnoses and their Characterization by Circumscription », *Annals of Mathematics and Artificial Intelligence 11*, 1994, p. 75–96.

[2] P. Boucher, P. Sébillot, « Interprétation et génération automatiques de noms composés anglais à l'aide de formes logiques », *Traitement Automatique des Langues 34*, 2, 1993, p. 89–104.

[3] M.-O. Cordier, P. Siégel, « Prioritized transitions for Updates », *in : Proceedings of European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, C. Froidevaux, J. Kohlas (éditeurs), *LNAI 946*, Springer, p. 142–151, 1995.

[4] P. Dupont, L. Miclet, E. Vidal, « What is the search space of the regular inference? », *in : Grammatical inference and applications, Lectures notes in Artificial Intelligence 862*, Springer-Verlag, p. 25–37, 1994.

[5] P. Dupont, L. Miclet, « Inférence grammaticale régulière : fondements théoriques et principaux algorithmes », *Rapport de Recherche n° 3449*, INRIA, Rennes, juillet 1998, également rapport IRISA 1189.

[6] I. Lerman, « Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème de consensus en Classification », *Revue de Statistique Appliquée XXXV*, 2, 1987, p. 39–60.

[7] I. Lerman, « Conception et analyse d'une famille de coefficients statistiques d'association entre variables relationnelles I, II », *Revue Mathématiques, Informatique et Sciences Humaines 30*, 118 et 119, 1992, p. 35–52, 75–100.

[8] Y. Moinard, R. Rolland, « Around a Powerful Property of Circumscriptions », *in : Actes de JELIA'94, LNCS No 838*, Springer-Verlag, p. 34–49, 1994.

[9] Y. Moinard, R. Rolland, « Preferential entailments for circumscriptions », *in : KR'94*, J. Doyle, E. Sandewall, P. Torasso (éditeurs), Morgan Kaufmann, p. 461–472, Bonn, mai 1994.

[10] Y. Moinard, R. Rolland, « Propositional circumscriptions », *rapport de recherche*, INRIA Research Report RR-3538, également Publication Interne IRISA 1211, Rennes, France, octobre 1998, `http://www.irisa.fr/EXTERNE/bibli/pi/1211/1211.html`.

[11] R. Pichon, P. Sébillot, « Acquisition automatique d'informations lexicales à partir de corpus : un bilan », *Rapport de Recherche n° RR-3321*, INRIA, décembre 1997.

[12] S. Thiébaux, M.-O. Cordier, O. Jehl, J.-P. Krivine, « Supply Restoration in Power Distribution Systems — A Case Study in Integrating Model-Based Diagnosis and Repair Planning », *in : Actes de UAI-96*, p. 525–532, 1996.

Thèses et habilitations à diriger des recherches

[13] J.-M. Guinnebault, *Caractérisation des logiques non-monotones et logique conditionnelle du deuxième ordre*, mémoire de doctorat, université de Rennes 1, mai 1999.

[14] E. MAYER, *Apprentissage inductif de scénarios pour la supervision de réseaux de télécommunications*, mémoire de doctorat, université de Rennes 1, décembre 1999.

[15] V. ROUAT, *Validité de l'approche classification dans la réduction statistique de la complexité de* #SAT, thèse de doctorat, université de Rennes 1, janvier 1999.

Articles et chapitres de livre

[16] C. BELLEANNÉE, P. BRISSET, O. RIDOUX, « A Pragmatic Reconstruction of Lambda-Prolog », *Journal of Logic Programming 41*, oct 1999, p. 67–102, `file:/udd/belleann/latex/papier/olivier/jlp-final.ps`.

[17] C. BELLEANNÉE, J. NICOLAS, R. VORC'H, *Le concept de preuve à la lumière de l'intelligence artificielle*, *Nouvelle Encyclopédie Diderot*, Presses Universitaires de France, nov 1999, ch. Vers un démonstrateur adaptatif.

[18] J.-M. BLIN, V. MASSON, R. QUINIOU, « Acquisition d'expérience par raisonnement à partir de cas dans un système de recherche », *Revue d'Intelligence Artificielle 13*, mar 1999, p. 73–95.

[19] C. DELAMARCHE, P. GUERDOUX-JAMET, R. GRAS, J. NICOLAS, « A symbolic-numeric approach to find patterns in genomes: Application to the translation initiation sites of *E. coli* », *Biochimie 81*, 1999, `nfs:/udd/jnicolas/windows/papiers/biochimie99.ps`.

[20] I.-C. LERMAN, V. ROUAT, « Segmentation de la sériation pour la résolution de #SAT », *Mathématiques, Informatique et Sciences Humaines*, 147, septembre 1999, p. 113–134.

[21] I.-C. LERMAN, « Comparing classification tree structures: A special case of comparing q-ary relations », *RAIRO Operations Research 33*, septembre 1999, p. 339–365.

Communications à des congrès, colloques, etc.

[22] P. BESNARD, M.-O. CORDIER, « Inferring causal explanations », *in : Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU'99)*, A. Hunter, S. Parsons (éditeurs), *Lecture Notes in Artificial Intelligence, 1638*, Springer-Verlag, p. 55–67, juillet 1999.

[23] G. CARRAULT, M.-O. CORDIER, R. QUINIOU, M. GARREAU, J.-J. BELLANGER, A. BARDOU, « A model-based approach for learning to identify cardiac arrhythmias », *in : Actes de AIMDM'99 : Artificial Intelligence in Medicine and Medical Decision Making*, W. Horn, Y. Shahar, G. Lindberg, S. Andreassen, J. Wyatt (éditeurs), *Lecture Notes in Artificial Intelligence, 1620*, Springer Verlag, p. 165–174, Aalborg, Denmark, jun 1999, `http://www.irisa.fr/aida/Pages_Pros/quiniou/iam/iam.pdf`.

[24] C. FABRE, P. SÉBILLOT, « Semantic Interpretation of Binominal Sequences and Information Retrieval », *in : Actes de CIMA'99 (International ICSC Congress on Computational Intelligence: Methods and Applications, Symposium on Advances in Intelligent Data Analysis AIDA'99)*, Rochester, N.Y., ÉTATS-UNIS, juin 1999.

[25] Y. MOINARD, R. ROLLAND, « À propos de la circonscription propositionnelle », *in : JNMR'99*, GDRI3, Paris, mars 1999, `http://www.irit.fr/GDRI3-ModRais/articlesJNMR.html`.

[26] A. OSMANI, L. ROZÉ, « Supervision of telecommunication networks », *in : European Control Conference*, septembre 1999.

[27] R. PICHON, P. SÉBILLOT, « Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience », *in : Actes de TALN'99 (Traitement automatique des langues naturelles)*, Cargèse, FRANCE, juillet 1999.

[28] R. PICHON, P. SÉBILLOT, « From Corpus to Lexicon: from Contexts to Semantic Features », *in : Actes de PALC'99 (International conference on Practical Applications in Language Corpora)*, Lodz, POLOGNE, avril 1999.

[29] B. Tallur, J. Nicolas, C. Delamarche, « Essai de prédiction de fonction de protéines de la famille MIP par la classification des séquences », *in : Les actes de SFC'99: Septièmes journées de la Société Francophone de la classification*, septembre 1999, `nfs:/udd/tallur/texte/papers/sfc/nancy99-short.ps`.

[30] B. Tallur, « A hierarchical clustering technique in biological sequence analysis », *in : Combinatorics, statistics, pattern recognition and related areas*, janvier 1999.

Rapports de recherche et publications internes

[31] F. Coste, « State Merging Inference of Finite State Classifiers », *Technical-report*, Irisa, mai 1999, `ftp://ftp.irisa.fr/techreports/1999/PI-1250.ps.gz`.

[32] Y. Moinard, R. Rolland, « Preferential entailments, extensions and reductions of the vocabulary », *rapport de recherche n° 3787*, INRIA, Rennes, France, octobre 1999, également rapport IRISA PI 1273, `http://www.irisa.fr/EXTERNE/bibli/pi/1273/1273.html`.

[33] J. Nicolas, « Grammatical inference as unification », *rapport de recherche n° 3632*, INRIA, juillet 1999, également rapport IRISA PI1265, `ftp://ftp.irisa.fr/techreports/1999/PI-1265.ps.gz`.

Divers

[34] P. Bouillon, C. Fabre, L. Jacqmin, P. Sébillot, « Rapport scientifique de fin de 1ère année, ARC Aupelf-Uref, Convention X/1.20.09.01.1/98.16.1 », décembre 1999.

[35] M.-O. Cordier, I. Grosclaude, R. Quiniou, « Amélioration des traitements des informations temporelles dans les graphes causaux temporels, Rapport de fin de contrat », jan 1999.