

*Projet LeD**Langue et Dialogue**Nancy*

THÈME 3A



*R*apport
d'Activité

2001

Table des matières

1	Composition de l'équipe	3
2	Présentation et objectifs généraux	4
3	Fondements scientifiques	5
3.1	Etude et modélisation de l'usage de la langue	5
3.2	Une approche centrée « référence »	6
4	Domaines d'applications	7
4.1	Panorama	7
4.2	Modèles informatiques et linguistiques du dialogue homme-machine	8
4.3	Systèmes de dialogue	9
4.4	Outils et méthodes pour la gestion et la manipulation de ressources linguistiques	9
4.5	Génération de textes et co-référence événementielle	10
4.6	Recherche dans les sources d'informations génomiques	10
5	Logiciels	11
5.1	Plateforme RM	11
5.2	Analyseur LTAG	12
5.3	Editeur graphique XML	12
5.4	Système EGAL	13
5.5	DHYDRO	13
5.6	Suite logicielle SALT	14
5.7	Alignement multilingue	15
5.8	XMAP	15
6	Résultats nouveaux	16
6.1	Analyse d'énoncés à l'aide de grammaires d'arbres lexicalisées	16
6.2	Métamodèle pour la terminologie multilingue	16
6.3	Domaines de référence	17
6.4	Sémantique lexicale	17
6.5	Aspects génération	18
6.6	Dialogue multimodal	19
6.7	Inférence dans le traitement automatique des langues	20
7	Contrats industriels (nationaux, européens et internationaux)	21
7.1	Thomson-LCR	21
7.2	EADS/MSI	21
7.3	Cinergie	21
7.4	Lucid-IT	22
7.5	Actions régionales	22
7.6	Actions nationales	23
7.7	Actions internationales	23

7.8	Actions de normalisation	24
7.9	Relations bilatérales internationales	25
8	Diffusion de résultats	26
8.1	Animation de la communauté scientifique	26
8.2	Enseignement universitaire	28
8.3	Autres enseignements	29
9	Bibliographie	29

Langue et Dialogue est un projet du LORIA (UMR 7503) commun au CNRS, à l'INRIA, à l'Université Henri Poincaré Nancy 1, à l'Université Nancy 2 et à l'Institut National Polytechnique de Lorraine.

1 Composition de l'équipe

Responsable scientifique

Laurent Romary [directeur de recherche Inria]

Responsable permanent

Jean-Marie Pierrel [professeur à l'UHP - Nancy 1]

Assistante de projet

Isabelle Blanchard [technicienne CNRS (à mi-temps)]

Personnel INRIA

Patrick Blackburn [directeur de recherche INRIA]

Laurence Danlos [professeur à l'Université de Paris 7, en délégation à l'INRIA-Lorraine jusqu'au 31/08/2001 puis associée]

Personnel CNRS

Marie-Dominique Devignes [chargée de recherche CNRS (département SDV, en stage au Loria)]

Bertrand Gaiffe [chargé de recherche CNRS]

Claire Gardent [chargée de recherche CNRS (département SHS)]

Nadia Viscogliosi [Ingénieur d'étude CNRS]

Personnel universitaire

Nadia Bellalem [maître de conférences - IUT Nancy-Charlemagne - Université Nancy 2]

Daniel Coulon [professeur - Ecole des Mines - INPL]

Samuel Cruz-Lara [maître de conférences - IUT Nancy-Charlemagne - Université Nancy 2]

Christine Fay-Varnier [maître de conférences - Ecole de Géologie-INPL]

Jean-Luc Husson [maître de conférences - IUT de Nancy-Brabois - UHP Nancy 1]

Azim Roussanaly [maître de conférences - UFR MI - Université Nancy 2 (en détachement à l'INRIA depuis 01/09/2001)]

Malika Smaïl [maître de conférences - UFR STMIA - UHP Nancy 1]

Chercheurs doctorants

Susanne Salmon-Alt [bourse MENRT - Monitrice - Université Nancy2 (jusqu'au 01/09/2001, puis associée/CR ATILF/CNRS)]

Evelyne Jacquy [ATER UHP - Nancy 1]

Frédéric Landragin [bourse MENRT]

Hélène Manuellian [bourse INRIA]

Djamé Seddah [bourse INRIA]

Ti Minh Nguyen [bourse du MAE et enseignante à l'Université d'Hanoi]

Dalila Beckhouche [bourse CIFRE EADS/MSI]

Benoît Crabbé [bourse INRIA/Région - Moniteur - Université Nancy2 (depuis 01/09/2001)]

Chercheurs associés et post-doc

Gérald Duffing [associé, enseignant à l'ICN, Université Nancy 2]
 Amalia Todirascu [post-doc depuis le 01/09/2001]
 Ruwan Weerasinghe [post-doc depuis le 01/09/2001]

Chercheurs invités

Nancy Ide [Vassar College, Ploughkeepsy (USA)]
 Maarten Maerx [Univ. Amsterdam (NL)]
 Milena Sclavcheva [Univ. Sc. Bulgarie (jusqu'au 30/11/2001)]
 Kristina Striegniz [Univ. Saarbruck (D)]

Ingénieurs experts

André Schaaff [Ingénieur expert sur contrat (projet HLT/SALT)]
 Eric Kow [Ingénieur expert depuis 25/10/2001]
 Annalisa Ricci [Ingénieur expert depuis 01/08/2001 (projet Cinergie)]

Autres personnels

Evelyne Bourion [Ingénieur d'étude à l'ATILF-CNRS préparant une thèse de linguistique sous la direction conjointe de François Rastier (ATILF-CNRS) et de Laurent Romary]
 Vincent Strohmenger [Stagiaire DESS en contrat d'apprentissage jusqu'au 01/08/2001]

2 Présentation et objectifs généraux

L'objectif du projet Langue et Dialogue est de définir des modèles et des techniques permettant de mettre en œuvre à court, moyen ou long terme des systèmes de dialogue homme machine finalisés reposant sur une forte composante langagière. Dans ce cadre, notre activité se développe dans trois directions complémentaires :

- l'étude des mécanismes fondamentaux de la communication en langue naturelle seule ou accompagnée d'une désignation gestuelle (dialogue multimodal). Cette recherche s'effectue dans un contexte pluri-disciplinaire alliant linguistique et informatique principalement ;
- la réalisation de systèmes de dialogue effectifs dans le cadre notamment de collaborations industrielles. Cette activité nous permet par ailleurs de disposer d'une plateforme d'expérimentation pour la validation des différents modèles que nous concevons ;
- la définition d'outils et de méthodes génériques permettant d'étudier de façon fine des situations de dialogues réels, issus de la transcription d'expériences de simulation ou d'observations directes. Ce travail s'appuie sur une expérience acquise depuis plusieurs années sur la normalisation et la manipulation de ressources linguistiques (en particulier des « corpus »).

Nous restreignons notre ambition à du DHM finalisé, c'est-à-dire dédié à la commande, à l'apprentissage ou à la recherche d'informations dans un domaine clairement spécifié ; cela suppose que l'on peut expliciter complètement l'action – ou plus généralement l'intention – visée par l'utilisateur, de manière à ce qu'elle puisse être réalisée par le système ¹.

Il ne s'agit, cependant, pas de se limiter à l'emploi de quelques bribes de langage naturel pour apporter une consonnance naturelle. Les recherches que nous menons, visent à long terme,

¹Dans la suite, nous parlerons d'application pour désigner ce domaine clairement spécifié

à mettre en œuvre une véritable communication en langue naturelle entre un usager et une machine.

Le choix de partir de l'hypothèse d'un dialogue oral — même si nous n'effectuons pas de recherche sur la parole proprement dite — résulte de la volonté de gérer au plus près la dynamique de la tâche et correspond à l'expérience acquise par les chercheurs du groupe dans la prise en compte des aspects temporels dans la langue ainsi que de phénomènes plus particuliers tels que la déixis, que l'on peut observer par exemple dans des expériences de simulation de dialogues oraux. De fait, il nous semble important, dans un souci de respect de la spontanéité des utilisateurs potentiels de nos systèmes, de savoir modéliser les modes de référencement immédiate aux différentes composantes d'une tâche donnée, notamment quand celle-ci est visualisée sur un écran.

3 Fondements scientifiques

3.1 Etude et modélisation de l'usage de la langue

Mots clés : langue, dialogue, syntaxe, sémantique, pragmatique.

Résumé :

La problématique de dialogue homme-machine qui est la nôtre nous conduit à nous placer plus dans la perspective de l'étude et de la modélisation de l'usage de la langue que dans celle de l'étude de l'objet « langue ». En ce sens on peut dire que notre approche est fondamentalement pragmatique, l'interprétation des énoncés jusqu'aux objets et le dialogue avec un utilisateur représentent nos points d'ancrage. Les modélisations sémantiques, syntaxiques et lexicales sont pour nous au service de ces objectifs premiers. Il en est de même des techniques de représentation des connaissances et d'Intelligence Artificielle que nous n'abordons pas comme objets d'étude mais plutôt comme moyens indispensables pour valider nos propositions plus théoriques sur l'usage de la langue en situation de dialogue homme-machine à forte composante langagière.

Lorsque la langue est le sujet d'étude, à tort ou à raison, l'un des enjeux théoriques est le découpage entre syntaxe, sémantique, etc. Nous nous plaçons, quant à nous, non pas dans la perspective de l'étude de l'objet « langue » mais dans celle de l'étude et de la modélisation de son usage. Tous les niveaux d'étude habituels (syntaxe, sémantique et pragmatique) sont cependant connexes; la différence essentielle avec les études descriptives est donc de définir celui que l'on souhaite privilégier, c'est-à-dire le type de résultats que l'on souhaite obtenir (la perfection exigerait que tous les résultats à tous les niveaux soient corrects mais la perfection est un projet à très long terme!). Pour notre part, nous avons choisi de nous concentrer sur l'usage de la langue : nous sommes donc dans une problématique d'interface et le résultat attendu est, en première analyse, un effet sur une application. En seconde analyse, cet objectif demeure mais s'insère dans une logique de dialogue; par conséquent, la commande de l'application reste le but mais suppose aussi un retour d'information correct et compréhensible par l'utilisateur humain. Comme indiqué ci-dessus, l'interprétation des énoncés jusqu'aux objets

et le dialogue avec un utilisateur représentent nos points d'ancrage. Dès lors, la sémantique dont nous avons besoin doit tout à la fois intégrer le dialogue et représenter l'application à commander. De même, la syntaxe n'est qu'au service de cette sémantique et du calcul de référence ; il est en particulier hors de question pour nous de considérer une syntaxe normative (i.e. qui contraindrait l'usage de la langue), ce qui reviendrait à être capable de corriger les fautes de français de l'utilisateur ; en revanche, et ce n'est pas nécessairement antagoniste, nous avons besoin d'une syntaxe capable de guider un système de reconnaissance de parole. Il suffit d'observer un corpus de parole avec ses faux départs, hésitations, etc. pour comprendre cette différence entre une syntaxe destinée à participer à l'interprétation d'un énoncé et une syntaxe destinée à engendrer des messages tels qu'on souhaite en lire. . . Un des problèmes auxquels nous devons faire face est le décalage entre, d'un côté, des mots de la langue liés directement ou non à des catégories d'objets, et la représentation informatique d'une application. Ce problème est évidemment lié à la représentation de connaissances. Pour autant, ce point ne nous intéresse :

- qu'en termes de méthodes de représentation de connaissances ; et sous cet aspect nous nous considérons plus comme consommateurs que comme producteurs ;
- qu'en ce qui concerne la représentation des objets et leur évolution temporelle et à partir de catégories liées à la langue. Cet aspect, au contraire, nous semble primordial pour l'objectif décrit.

Enfin, la mise en œuvre de systèmes de dialogues pose un double problème d'architecture :

- d'une part, il s'agit de gros programmes avec des problèmes de modularité et de contrôle liés à l'imbrication de plusieurs sources de connaissances ;
- d'autre part, il faut une réflexion intégrant à la fois des aspects liés au langage et des aspects liés à la perception par un utilisateur des effets de ses actions et des objets sur lesquels il peut agir. Cette réflexion amène inévitablement à des questions de modularité au sens cognitif.

Dans cette mesure, si les outils permettant la mise en œuvre de systèmes distribués ne sont pas pour nous un objet d'étude, leur utilisation et l'implémentation de systèmes de dialogues dans des outils préexistants nous semblent indispensables pour valider nos propositions plus théoriques. Quelques conséquences immédiates découlent de ce qui précède :

- nous considérons la mise en œuvre de systèmes de dialogue à la fois comme un outil de validation de nos propositions et comme la garantie que nous nous attaquons à des problèmes réels et dans un ordre réaliste quant à leur importance effective. . .
- la modélisation de l'état de l'application dans des termes compatibles avec un accès langagier (en particulier en ce qui concerne la représentation du temps) nous semble essentielle. . .
- l'historique du dialogue, pendant langagier de l'état de l'application, est l'autre préalable à toute mise en œuvre de système de dialogue.

3.2 Une approche centrée « référence »

Mots clés : référence, analyseur local, multimodalité, geste, corpus.

Résumé :

Conséquence directe du point 3.1., notre approche est une approche centrée

langue et référence qui nous conduit tout à la fois à une vision très locale de l'analyse syntaxique, très dépendante du dialogue et du contexte d'application pour la sémantique, très liée à leur contexte langagier d'utilisation pour les gestes de désignation. Par ailleurs, afin de coller à un usage véritable de la langue, l'exploration de corpus s'avère pour nous indispensable à la fois comme source de réflexion et comme moyen de validation.

Les points précédents motivent à la fois notre travail actuel et notre orientation pour l'avenir. Le travail sur l'évolution de l'application nous a amenés en particulier à définir un modèle temporel original et la réflexion sur l'historique a motivé nos propositions sur la référence. L'un de nos objectifs pour la suite vise à intégrer une partie même de l'historique du dialogue dans la représentation des objets. De ce parti pris de nous centrer sur la référence et (donc) le dialogue découlent :

- une vision de la syntaxe conduisant vers des analyseurs très locaux... S'il s'agit d'une réponse à une question portant sur la couleur d'un objet, « Rouge! » est un énoncé parfaitement admissible ;
- une vision de la sémantique subordonnée elle aussi au dialogue. A titre d'exemple, la présence obligatoire d'un objet direct pour un verbe transitif se pose en des termes différents selon qu'on considère des phrases dans un texte, des énoncés isolés ou des séquences d'énoncés dans des dialogues oraux ;
- une analyse des gestes de désignation (recueillis à la souris, au gant de désignation ou à l'aide d'un écran tactile suivant les cas) intrinsèquement liée à leur utilisation dans des énoncés. Typiquement, dans cette optique, le même geste se verra attribuer une interprétation ou même une structure différente selon l'énoncé langagier par rapport auquel il s'interprète.

En termes d'outils, si l'implémentation d'un système de dialogue est une source de réflexion pour toute une équipe (au cours même de l'implémentation, mais surtout après, lorsqu'on évalue son caractère plus ou moins approprié au problème de départ), l'exploration de corpus s'avère également précieuse et surtout plus rapide pour autant qu'on dispose d'outils associés à la manipulation de ces corpus. En particulier, les corpus obtenus par la technique du magicien d'Oz² permettent de se projeter au-delà des capacités réelles des systèmes de dialogue existants (le cas le plus évident est celui de la reconnaissance de parole, mais cela peut également être vrai de tout autre « module »). Ils sont donc pour nous une source de réflexion et un moyen de validation et non pas par exemple un moyen de faire apprendre des modèles statistiques.

4 Domaines d'applications

4.1 Panorama

Les recherches que nous menons au sein du projet Langue et Dialogue se déclinent en trois points :

²Il s'agit d'expériences de simulation au cours desquelles le système de dialogue est remplacé par un expérimentateur humain qui interagit avec l'utilisateur et la tâche.

- des études pluridisciplinaires sur la langue : comme nous l'avons vu, l'objectif est de mettre l'accent sur les conditions d'usage de la langue en combinant observation et réflexion linguistique d'une part et modélisation informatique d'autre part ;
- la réalisation de prototypes : ceci nous permet de valider nos modèles dans le cadre de situations concrètes, inspirées notamment d'applications industrielles réalistes ; c'est dans ce cadre que se situent notamment les travaux menés sur la recherche d'information en génomique décrits spécifiquement dans la section 4.6 ;
- la définition d'outils et méthodes pour l'étude de corpus : cet aspect est fondamental pour fournir à la fois une base solide pour les études linguistiques que nous menons et faciliter le test de méthodes particulières (par exemple de l'analyse syntaxique locale ou d'une stratégie de calcul référentiel) sur des données réelles ;

4.2 Modèles informatiques et linguistiques du dialogue homme-machine

L'une des caractéristiques des recherches que nous menons est qu'il est impossible d'isoler une composante spécifique du traitement du langage. A ce titre, tout élément d'analyse, qu'il porte sur le niveau lexical, syntaxique ou sémantique doit s'intégrer dans une vision plus large permettant *in fine* l'interprétation de l'énoncé courant de l'utilisateur et la gestion des réponses qui lui sont fournies (lancement d'une action, message de réponse, etc.). Depuis plusieurs années, notre projet a placé le phénomène de la référence au sens large au centre de cette notion d'interprétation. Ceci nous a en particulier conduits à proposer un modèle original centré sur la notion de contraste (toute opération de référence isole une entité qui peut être distinguée d'autres entités similaires). Nos efforts se concentrent plus particulièrement sur les aspects suivants :

- définition d'un modèle de représentation des référents : les Représentations Mentales (RM) – l'objectif de cette recherche, menée plus particulièrement dans le cadre du projet Cervical, est d'intégrer, dans la mesure du possible, l'ensemble des informations attachées à un objet susceptible d'être activé lors d'un acte de référence. En particulier, notre objectif est d'expliquer comment, dans une même représentation, peuvent se combiner des informations d'ordre spatial et/ou perceptif, temporel et lexical notamment. Certains de ces points font l'objet d'études spécifiques ;
- lien entre geste et perception dans le dialogue multimodal — suite aux travaux de thèse de Nadia Bellalem et de Frédéric Wolff, nous avons mené une étude expérimentale sur la relation entre les trajectoires gestuelles d'un utilisateur et la structure perceptive des objets présentés à lui. L'analyse du corpus que nous avons recueilli devrait nous permettre de définir des algorithmes plus fins d'analyse des trajectoires référentielles ;
- de nouvelles expérimentations ayant pour but de relier ces résultats aux travaux de Laurent Romary sur la référence sont en cours dans le cadre des thèses de Susanne Salmon-Alt et Frédéric Landragin
- temps et prédication dans le dialogue — dans le cadre de la collaboration avec l'université de Genève (accueil de deux doctorants de l'équipe de Jacques Moeschler - Département de linguistique - Université de Genève), nous avons repris le travail de modélisation des informations temporelles dans le dialogue, notamment pour mieux représenter les actions élémentaires relativement à un référent ou un groupe de référents donné ;

- gestion du dialogue – les travaux de thèse d'Olivier Grisvard nous ont menés à proposer une représentation des énoncés intégrant force illocutoire et contenu propositionnel en termes d'événements inclus dans le contexte. L'assemblage de sommes de tels événements mène à une modélisation du dialogue sur laquelle le travail se poursuit actuellement ;
- analyseurs TAG pour le dialogue — nous avons adapté le modèle syntaxique des TAG (Tree Adjoining Grammar) pour lui permettre de prendre en compte des données issues de systèmes de reconnaissance de la parole, par le biais de stratégies d'analyse locales par connexité. Le travail de thèse de Djamé Seddah s'inscrit dans ce cadre et a pour objectif d'obtenir une véritable forme logique à l'issue de l'analyse.

4.3 Systèmes de dialogue

Dans le cadre du projet Esprit II Multiworks, nous avons eu l'occasion de développer une première plateforme de dialogue reposant sur les différentes propositions théoriques (modèle temporel à deux relations, première version du modèle référentiel) faites au sein de l'équipe. Nous avons fait le choix de faire évoluer cette plateforme (« Multidial ») pour qu'elle puisse à la fois servir de site d'expérimentation pour nos propositions théoriques et de base pour les applications futures, dans le cadre notamment de nos collaborations industrielles. Notre objectif est de développer plus particulièrement les aspects suivants :

- définition d'un module lexical plus flexible – afin de palier les difficultés liées à la re-définition d'une composante lexicale pour chaque nouvelle application, nous étudions actuellement l'usage qui peut être fait de ressources normalisées issues notamment des projets européens Eagles et Multext. En particulier, nous étudions comment il est possible de « surcharger » des informations morphosyntaxiques plus ou moins standard par des indications sémantico-pragmatiques plus spécifiquement dédiées à la tâche en cours ;
- remplacement de l'analyseur existant par un analyseur TAG – en parallèle avec les travaux plus théoriques que nous menons sur ce sujet, nous développons un analyseur TAG reposant sur des algorithmes d'analyse classiques ^[Ear] ;
- évaluation d'une architecture distribuée – nous menons actuellement différentes expérimentations pour évaluer l'intérêt d'une norme telle que Corba pour la réalisation d'une architecture de système de dialogue distribuée.

Ces recherches sont en particulier développées dans le cadre de la collaboration que nous entretenons avec Alcatel Business Systems à Strasbourg et avec qui nous participons au projet Eureka Master.

linguistiques

4.4 Outils et méthodes pour la gestion et la manipulation de ressources linguistiques

Gérer des ressources linguistiques normalisées représente pour nous à la fois un moyen de valider, par l'observation, nos modèles théoriques et une source générique d'information (par exemple lexicale) pour les systèmes de dialogue que nous développons. Plus globalement, il s'avère que la réflexion menée autour de la normalisation des ressources linguistiques doit

[Ear] C. EARLEY, JAY, « An Efficient Context-free Parsing Algorithm », *Communication of ACM*.

être largement partagée au sein de la communauté académique. C'est à ce titre qu'il nous a été demandé de coordonner le projet Silfide dans le cadre duquel nous situons l'essentiel de nos actions. Le projet Langue et Dialogue occupe ainsi une place active au sein de la communauté nationale et internationale dans le domaine de la normalisation des ressources linguistiques et de leur utilisation, notamment au sein de la TEI (Text Encoding Initiative <http://www.tei-c.org>). Parmi les points que nous développons plus particulièrement, nous pouvons mentionner :

- le développement de nos compétences dans les domaines des lexiques et des dictionnaires. Nous bénéficions en particulier de l'expérience acquise au sein du projet MLIS/Dhydro ;
- la définition de standards pour l'annotation de corpus de dialogues – c'est ainsi que nous avons associé aux travaux sur l'analyse à base de grammaires TAG, un schéma d'annotation (TAGML) reposant sur XML ;
- le renforcement de nos activités dans le domaine du traitement multilingue. Une collaboration est ainsi en cours avec l'université de Hanoï (thèse de Ti Min Huyen en co-tutelle) sur la définition d'outils communs pour le Français et le Vietnamien.

4.5 Génération de textes et co-référence événementielle

Grâce à l'accueil en délégation INRIA de Laurence Danlos, nous avons initié, au sein de l'équipe, des recherches sur la génération de textes qui s'appuient sur G-TAG, un formalisme pour générer des textes à partir de leur représentation conceptuelle. Ce formalisme est conçu pour utiliser les informations syntaxiques et lexicales d'une grammaire lexicalisée d'arbres adjoints (notée LTAG ou simplement TAG) pour une langue cible. Cette grammaire TAG est étendue de façon à couvrir des textes de plusieurs phrases. Elle est aussi complétée par une interface conceptuel-sémantique. Cette interface est lexicalisée : c'est un ensemble de bases de données, comme c'est le cas pour l'interface sémantique-syntaxe, i.e. la grammaire TAG. De ce fait, G-TAG est un formalisme lexicalisé pour la génération de textes, ce qui est une approche novatrice dans la mesure où la lexicalisation en génération n'avait été envisagée que pour la production de phrases. Ce formalisme est décrit dans ^[Dan98], ses applications dans ^[MD98].

Le travail de thèse d'Hélène Manuélian se fonde sur ce formalisme et porte plus particulièrement sur la génération d'expressions référentielles (cf. 6.5).

4.6 Recherche dans les sources d'informations génomiques

La recherche d'informations sur les génomes nécessite encore souvent d'effectuer des requêtes multiples à des bases de données hétérogènes accessibles via Internet. Les différences rencontrées dans les formats des données, les problèmes de nomenclature (alias, synonymes) et le manque de documentation concernant la validité des données constituent des obstacles certains pour l'unification des données recueillies et la production d'une information utilisable.

[Dan98] L. DANLOS, « G-TAG : un formalisme lexicalisé pour la génération de textes inspiré de TAG », *Traitement Automatique des Langues - T.A.L.* 39, 2, 1998.

[MD98] F. MEUNIER, L. DANLOS, « FLAUBERT: an User friendly system for Multilingual Text Generation », in : *Actes 9th International Workshop on Natural Language Generation (INLG'98)*, p. 284–287, Niagara-on-the-Lake, 1998.

Une solution à ce problème consiste à améliorer la qualité des bases de données biologiques en termes d'intégration des données et d'unification de leur représentation, de mise à jour, de documentation, et aussi à nettoyer les données à travers des procédures de confrontation et de validation. Diverses initiatives ont eu pour but de proposer à l'utilisateur de nouvelles bases de données plus « propres ». Or malgré la qualité certaine de ces ressources supplémentaires, celles-ci ne répondent pas toujours aux besoins du scientifique qui souhaite compiler toutes les données disponibles sur un ou plusieurs objets donnés (gène, marqueur, séquence, etc.). L'objet d'intérêt peut tout simplement être absent des ressources proposées ; de plus, des informations complémentaires peuvent être présentes dans des sources non prises en compte par ces ressources mais utiles à consulter pour tel ou tel type d'objet. L'utilisateur peut alors bénéficier de systèmes d'interrogation intégrés tels que Entrez au NCBI (<http://www.ncbi.nlm.nih.gov/Entrez>) ou SRS à l'EBI (<http://www.ebi.ac.uk/srs>). La limite majeure de ces systèmes concerne la gestion des réponses aux requêtes, surtout lorsqu'elles sont multiples et concernent des bases de données hétérogènes. De plus, les systèmes restent neutres par rapport à la qualité des bases de données interrogées et à la fiabilité des réponses obtenues.

Dans un premier temps, nous nous sommes intéressés à ce problème en nous plaçant du point de vue de l'utilisateur et nous avons proposé un modèle de recherche d'information dans des bases de données hétérogènes accessibles sur le Web qui implique une assistance à la navigation et la structuration des données récoltées dans le but de faciliter leur exploitation ultérieure. Une DTD (Définition de Type de Document) spécifique a été conçue pour structurer et stocker les données collectées pendant une session de recherche dans un document conforme au standard XML. Le prototype Xmap a ainsi été développé [11] (cf. §5.8).

Cette première étude nous aura permis de nous familiariser avec les caractéristiques des bases de données biologiques utilisées en Génomique. Les limites que nous nous étions fixées (approche orientée utilisateur, DTD spécifique du problème de cartographie posé, interaction obligatoire avec l'utilisateur à de nombreuses étapes) doivent maintenant être dépassées et nous souhaitons aller plus loin dans l'interopérabilité des bases de données et élargir la palette de problèmes biologiques traités. C'est ainsi que nous travaillons en collaboration avec Nacer Boudjlida de l'équipe ECOO sur un environnement pour la génomique ouvert, distribué et orienté services.

5 Logiciels

5.1 Plateforme RM

Dans le cadre du post-doc industriel d'O. Grisvard, une implantation de la théorie des représentations mentales (RM) a été réalisée en partenariat avec l'équipe Communication Homme-Machine du Laboratoire Central de Recherches de Thalès. L'équipe Langue et Dialogue a fourni les algorithmes et les données théoriques sur lesquels cette implantation repose et Thalès la plateforme logicielle et le cadre applicatif dans lequel elle s'insère. Les RM s'intègrent à un système de dialogue oral homme-machine multi-agents sous la forme d'agents, chargés dans un premier temps de l'interprétation pragmatique des énoncés et à plus long terme, de la gestion du dialogue. L'objectif premier de cette implantation était de montrer que la théorie des RM est un modèle opératoire et efficace du calcul de la référence, objectif d'ores et déjà atteint

puisque l'agent chargé de la résolution de la référence est opérationnel. Par ailleurs, ont été implémentés le traitement de la partie prédicative des énoncés et la production des énoncés du système. Au sein de l'équipe, cette plate-forme sert actuellement de base pour la validation des algorithmes du calcul référentiel proposés par Susanne Salmon-Alt. Par ailleurs, nous envisageons son intégration dans une nouvelle plateforme de dialogue homme-machine et étudions actuellement les questions d'interfaçage avec l'analyseur syntaxique LTAG, disponible dans l'équipe.

5.2 Analyseur LTAG

Le travail sur l'analyseur syntaxique à base de LTAG élaboré par Patrice Lopez dans le cadre de sa thèse [5] s'est poursuivi cette année encore. Désormais, le format d'entrée des grammaires se conforme au standard TagML décrit en commun avec les équipes ATOLL de l'INRIA Rocquencourt, le DFKI/Sarrebrück et TALANA/Paris7[1]. Le travail se poursuit, en particulier grâce au travail d'Amalia Todirascu pour permettre à l'analyseur d'accepter un automate (plutôt qu'une chaîne de mots) en entrée, ce qui permettra d'utiliser plus facilement l'analyseur en sortie d'un système de reconnaissance de parole.

Les autres améliorations en cours de cet analyseur concernent la vérification des structures de traits durant l'analyse et la production des résultats sous forme de grammaires hors-contexte (décrites également dans le cadre de TagML) afin d'en faciliter l'utilisation dans des modules ultérieurs d'analyse (calcul sémantique et référentiel).

5.3 Editeur graphique XML

Nos activités dans le domaine du traitement de la langue naturelle, liées à l'effort que nous faisons de nous conformer à des standards reconnus de codage XML nous amènent à éditer et visualiser de nombreux documents XML.

Certains de ces documents, au delà de leur format XML d'échange avec d'autres partenaires, ont pour leurs utilisateurs une représentation graphique naturelle. On peut penser en particulier aux arbres (arbres élémentaires, arbres dérivés et arbres de dérivation d'une grammaire TAG), et aux structures de boîtes (représentations en sémantique dynamique type DRT, ou structures de traits lexicales).

Ceci a motivé le travail de Christophe Cocquet pour développer un éditeur graphique de documents XML (XML Editor disponible à <http://www.loria.fr/equipes/led/outils/xmlEditor/index.html>) Cet éditeur va bien au delà des produits existants (qui ne font que visualiser le document XML sous une forme embellie) en ce qu'il permet de travailler véritablement sur la visualisation des documents.

Par ailleurs, si pour certains utilisateurs l'éditeur se présente directement sous la forme d'un éditeur d'arbres ou de structures de traits, il est paramétrable à tout type de document XML et de visualisation pour autant qu'il soit paramétré par un fichier de transformation XSL. Notons enfin, que le format graphique obtenu finalement est SVG (Scalable Vector Graphics), c'est à dire un format vectoriel d'une part, et codé en XML d'autre part.

5.4 Système EGAL

Le système EGAL est une plateforme d'extraction de grammaire d'arbres adjoints lexicalisée. Elle intègre notamment l'analyseur syntaxique cité plus haut et un module interactif de construction d'un lexique syntaxique à partir d'un corpus.

Ce dernier a subi, cette année, une amélioration grâce :

- au développement d'un compilateur acceptant le nouveau formalisme proposé dans [59],
- à la mise en place d'un éditeur pour la visualisation et la maintenance des tests linguistiques,
- et à la mise à jour de l'interface du module.

5.5 DHYDRO

Le projet Dhydro contracté dans le cadre du programme européen MLIS s'est achevé le 26 mai 2000 par une présentation officielle des résultats à Bruxelles. Ce projet a permis de mettre au point un environnement de gestion collegiale de bases terminologiques multilingues, contenant plusieurs outils :

- Un outil de rétroconversion de 3 dictionnaires monolingues polysémiques codés selon le chapitre 12 de la TEI en une base conceptuelle trilingue codée selon la norme MARTIF ISO 12200.
- Un outil d'édition de fiches terminologiques multilingues, permettant l'importation de données depuis le serveur puis une exportation des fiches modifiées. Cette application comporte notamment une aide hypertextuelle trilingue et un outil de navigation dans une base de textes trilingues de référence, alignés par paire de langues.
- Un serveur permettant de gérer les bases terminologiques centralisées ainsi que les transactions issues des clients, d'administrer le système (gestion de comptes utilisateurs, de leurs droits et niveaux de responsabilités).
- Un outil d'indexation des bases terminologiques (complément de l'application serveur) permettant de répondre efficacement aux requêtes de consultation des données et de présenter les résultats sous divers formats grâce au mécanisme de feuilles de style XSL.
- Une interface web de consultation de la base terminologique accessible au grand public.
- Un outil d'administration permettant un suivi des transactions opérées sur le serveur.
- Un site éditorial web offrant un accès direct aux outils précédents.

Une version de ce système est actuellement opérationnelle au centre technique de la marine nationale a Brest (EPSHOM) pour permettre a plusieurs experts internationaux d'enrichir une base conceptuelle dans le domaine de d'hydrographie (environ 6000 concepts instanciés de manière trilingue en anglais, français et espagnol).

La plateforme Dhydro a fait l'objet de démonstrations régulières lors de diverses manifestations scientifiques et un certain nombre de demandes d'adaptation de ce système à leurs propres besoins ont émané d'organismes et reseaux internationaux.

- Organisation Maritime Internationale
- Réseau africain de terminologie RIFAL
- Réseau ibérique de terminologie RITERM
- Groupe EADS

Les conditions permettant d'exploiter Dhydro avec ces nouveaux partenaires sont actuellement à l'étude (Afnor, Office international des épizooties). De surcroît, des activités de formation à ce système sont aujourd'hui en cours dans plusieurs pays d'Afrique. Conjointement, l'institut de traducteurs et d'interprètes ISTI (Bruxelles) a jugé Dhydro particulièrement pédagogique et a par conséquent planifié une formation systématique de leurs étudiants à ce système.

Dhydro a été déposé officiellement comme logiciel INRIA et a fait l'objet d'un transfert industriel auprès de la société Lucid-IT (<http://www.lucid-it.com>). Cette plate-forme y sera exploitée pour développer un outil de développement collégial de bases terminologiques multilingues intégré à un environnement de gestion globale de la localisation de documents électroniques.

5.6 Suite logicielle SALT

Jusqu'à présent, plusieurs projets (GENERATER, Eurodicautom, OLIF, etc..) ont abouti, dans le domaine de la terminologie, à la création de formats de données mais sans grande concertation. Le but principal du projet SALT est donc de créer un format pivot permettant de faciliter les échanges et de réaliser les outils nécessaires (éditeur, routines de conversion, etc.) à l'exploitation de ce format pivot <http://www.loria.fr/projets/SALT/>. Les différents formats existants reposent tous sur un même principe : un même schéma conceptuel, la présence de catégories de données ainsi qu'un certain nombre de règles spécifiques définissant les manipulations de ces catégories. Ce format pivot, appelé GMT, est en cours de description dans le cadre de la création de la norme ISO16642 sous la responsabilité rédactionnelle de Laurent Romary. Un réel travail d'implémentation accompagne la création de cette norme.

Cette implémentation se traduit par la mise en œuvre d'une suite d'outils d'édition/manipulation de catégories de données et de conversion inter-formats appelée " Salt Suite ".

Celle-ci doit permettre aux utilisateurs de manipuler facilement des catégories de données (création de nouvelles catégories, extension ou restriction de catégories existantes par rapport à un ensemble de catégories issues d'une norme particulière (exemple : ISO12620)).

Par ailleurs, celle-ci doit également permettre de générer automatiquement une partie des filtres de conversion permettant de passer d'un format particulier vers le format pivot puis de passer du format pivot vers le format cible et également de faire cela dans l'autre sens dans le cadre d'une rétro-conversion.

Les développements sont intégralement réalisés en langage Java. Par ailleurs nous avons recours à d'autres technologies telles que XML (eXtensible Markup Language), XSL (eXtensible Stylesheet Language) et RDF (Ressource Description Framework). XML est un méta-langage de définition de langages de balisage et est actuellement au cœur des travaux du World Wide Web Consortium (W3C) et de nombreuses spécifications (à l'état de recommandation ou en cours) sont basées sur celui-ci.

Nous utilisons également le langage de spécification UML. RDF est une spécification pour la gestion des ressources et nous l'utilisons comme format de représentation des catégories de données. XSL est une spécification permettant de décrire des feuilles de style pour les documents conformes au standard XML. Nous utilisons en réalité XSLT qui est la partie transformation de XSL. Elle est utilisée principalement pour la partie conversion. En effet, les membres du

projet étant, pour moitié environ, linguistes et non informaticiens, il est impératif de rendre certaines parties accessibles à ces personnes. Par ailleurs, cette démarche offre également une grande souplesse au niveau de l'exploitation et de la dissémination des routines de conversion. La rapide évolution des technologies gravitant autour de XML nous oblige à être très réactif pour pouvoir intégrer rapidement les derniers changements validés par le W3C. Les outils développés sont destinés à être accessibles gratuitement pour tous les professionnels (et notamment de grandes entreprises) du domaine. Par conséquent, ces outils ne sont pas seulement des prototypes (ce qui en soi n'est pas péjoratif) mais sont susceptibles d'être appréhendés comme outils de référence par de nombreux utilisateurs potentiels.

5.7 Alignement multilingue

L'alignement multilingue de textes, c'est-à-dire la mise en correspondance à un certain niveau de granularité (paragraphe, phrase ou expression) d'un texte et de sa traduction dans une autre langue, est une des composantes essentielles des recherches effectuées par de nombreuses équipes en matière d'extraction d'informations multilingues et, d'un point de vue industriel, dans le domaine de la localisation. Les travaux menés dans l'équipe depuis plusieurs années [8] ont conduit au développement d'un module d'alignement (développé en Java) reposant sur une technique utilisant l'organisation hiérarchique de textes codés en divisions (récursives), paragraphes et phrases. Ce module a été testé sur de nombreuses paires de langues (notamment européennes) dans le cadre du projet Telri et est en cours de validation sur un corpus de textes franco-vietnamiens en cours de constitution. Complètement documenté dans le cadre du projet DHYDRO, et accompagné d'un concordancier multilingue permettant de filtrer les textes en fonction de contraintes exprimées sur des contextes linguistiques propres aux langues source et cible considérées, il est maintenant distribué librement (<http://www.loria.fr/equipes/led>).

5.8 XMAP

Le logiciel Xmap (<http://www.loria.fr/projets/Xmap/Index.htm>) a été développé pour la recherche de données de cartographie sur le génome humain ainsi que la corrélation de la position de nouveaux gènes avec celle de pathologies orphelines [11]. Ce logiciel comporte trois modules :

- *Xmap_retrieve* est un assistant de navigation pour structurer, collecter et sauvegarder des informations génomiques dans le format XML ;
- *Xmap_show* est un outil de visualisation intégrée des données recueillies par le module précédent ;
- *Xmap_multimodal* permet une interaction par la voix et le geste avec l'outil de visualisation *Xmap_show*.

Le prototype est actuellement en test dans le laboratoire de Charles Auffray, au CNRS à Villejuif. La version actuelle du logiciel, développé en Java, fonctionne sur machine Unix (SUN solaris), Linux et Windows (98 et NT). D'autres développements continuent de se faire : mise en place d'un manuel d'utilisation en ligne, analyse des mémoires de sessions afin de répondre à deux types de besoins, d'une part, l'optimisation du système en fonction de l'expérience acquise au cours des sessions, et d'autre part, la mise en place d'un système d'indexation

des sessions pour permettre aux utilisateurs de retrouver rapidement des données issues des sessions antérieures.

6 Résultats nouveaux

6.1 Analyse d'énoncés à l'aide de grammaires d'arbres lexicalisées

Cette année, s'est mise en place l'ARC RLT (Ressources Linguistiques pour les TAG) en collaboration avec les équipes ATOLL/INRIA Rocquencourt, CALLIGRAMME/LORIA TALANA/Paris7. C'est dans ce cadre que s'est déroulé le travail de DEA de Benoît Crabbé [59] qui a permis de faire le lien entre les sources de connaissances utilisées dans la plateforme EGAL [5] et la Métagrammaire (MG) développée à TALANA [Can99].

Schématiquement, on peut considérer que la MG permet d'engendrer automatiquement un ensemble de schèmes (arbres TAG qui ne sont pas reliés à un lexique), tandis que, de manière complémentaire, EGAL vise à relier les lexèmes à des ensembles de schèmes. Afin de garantir une cohérence entre ces deux processus, nous avons proposé un formalisme qui permet de décrire l'ensemble des propriétés syntaxiques utilisées dans la MG et de définir un ordre correspondant à des séquences de tests linguistiques. La compilation des connaissances exprimées dans ce formalisme permet de produire un arbre de décision qui est destiné actuellement à être intégré dans EGAL pour traiter un sous-langage d'application quelconque (caractérisé par une faible couverture lexicale). Dans la perspective d'un travail de thèse, nous envisageons de traiter des corpus à plus large couverture en nous appuyant sur un étiquetage automatique et une récupération manuelle des erreurs dans EGAL.

6.2 Métamodèle pour la terminologie multilingue

L'objectif de ce travail, conduit dans le cadre du projet européen HLT/Salt et qui forme la base de la proposition de future norme ISO 16642, consiste à définir un format générique de représentation de données terminologiques multilingues (TMF - Terminological Markup Framework). Il s'agit de décrire toute base de données terminologique comme la combinaison d'un squelette structurel représentant l'organisation hiérarchique de telles données (Collection d'entrées, entrée terminologique, informations propres à une langue, informations propres à un terme, informations propres aux composantes d'un terme) et d'un ensemble de caractéristiques décrites sous la forme de traits atomiques ou complexes (e.g. définition, liens conceptuels, information grammaticales, etc.).

A partir de la description de ces composantes, on peut spécifier intégralement toute représentation XML conforme à ces données générique et générer automatiquement, d'une part, les schémas XML permettant de contrôler ces formats, et d'autre part, les filtres permettant de transformer ces formats vers un format pivot XTMF (et réciproquement).

Ce travail préfigure des modèles génériques pour la représentation de tout type de données linguistiques (cf. collaboration avec Nancy Ide au Vassar College).

[Can99] M.-H. CANDITO, *Structuration d'une grammaire LTAG : application au français et à l'italien*, thèse de doctorat, Thèse de l'Université de Paris 7, 1999.

6.3 Domaines de référence

La thèse de Susanne Salmon-Alt [15] propose une modélisation d'une partie des mécanismes d'interprétation d'un énoncé en dialogue homme-machine : la résolution de la référence aux objets. Les approches courantes pour effectuer l'interprétation automatique d'une expression référentielle consistent essentiellement à filtrer un ensemble d'entités contextuelles en fonction des informations linguistiques et, éventuellement, des gestes de désignation associés. Le problème principal est alors de déterminer l'extension, la structuration et les mécanismes de mise à jour dynamique de cet espace de recherche. Une étude des théories et systèmes existants montre que les solutions mises en {oeuvre varient considérablement selon le type des expressions traitées (pronom vs. description définie), l'application envisagée (renseignement vs. commande), l'environnement d'interaction (langagier vs. multimodal) et la réversibilité ou non du système pour la génération. Or, une observation des données linguistiques dans des conditions naturelles montre qu'un dialogue réunit toutes ces caractéristiques à la fois. Partant de l'hypothèse qu'une trop grande diversité de mécanismes n'est ni plausible d'un point de vue cognitif, ni souhaitable d'un point de vue informatique, nous nous sommes attachée à proposer un modèle qui permette d'intégrer différents aspects du calcul référentiel dans un cadre théorique unifié [57]. Nous proposons une modélisation du contexte par des domaines de référence qui sont des sous-ensembles d'entités discursives ou perceptives, structurées selon des contrastes. Ensuite, nous faisons l'hypothèse que différents types d'expressions référentielles imposent différentes contraintes sur l'extraction de leur référent de ce contexte, et qu'elles le mettent à jour par différentes opérations de restructuration.

A partir de là, nous sommes capable de formuler des prédictions portant à la fois sur l'acceptabilité de telle ou telle expression dans un contexte donné et sur des interprétations préférentielles en cas d'ambiguïté. La portée théorique du modèle à été validée, pour l'interprétation des expressions en autre, sur un corpus de dialogues homme-homme [56]. Au passage, nous avons élaboré une méthodologie d'annotation de corpus qui dépasse les limitations imposées jusqu'alors par le cantonnement à la notion de coréférence, notoirement insuffisante pour tenir compte du statut référentiel de «autre» [32].

Enfin, l'implémentation qui a été faite de notre modèle montre que nos propositions peuvent être intégrées dans une application de véritable dialogue homme-machine.

6.4 Sémantique lexicale

L'un des enjeux importants de notre équipe est de développer des systèmes qui soient valides aussi bien sur le plan linguistique que sur le plan opératoire. Dans ce cadre, les phénomènes d'ambiguïté constituent une voie de recherche importante. Parmi les nombreuses sources d'ambiguïtés, un type particulier de polysémie nominale a été étudié : la polysémie logique (des noms d'objets comme "livre", objet physique ou informationnel, ou des noms d'action comme "acquisition", action ou résultat de l'action).

Dans le domaine de l'ingénierie des langues, la plupart des systèmes contournent la difficulté par le biais d'une modélisation éclatée : plusieurs entrées lexicales et des mécanismes de sélection contextuelle. L'une des conséquences importantes de cette approche est la perte de réutilisabilité d'un système de dialogue à l'autre.

Notre objectif, en particulier dans le cadre de la thèse d'Evelyne Jacquey, est donc de développer un procédé automatique de la gestion de ce type de polysémie qui soit compatible avec différents types d'application et en particulier avec différents points de vue possibles sur la caractérisation des entités d'une application désignées par des noms polysémiques.

S'appuyant sur les travaux actuels d'inspiration linguistique, [Pus95], [GJ96], [Kle99], [Cru95] et sur des travaux dédiés à la manipulation de types complexes dans le domaine de la sémantique formelle [AP00] et [PK00], la thèse d'Evelyne Jacquey a permis de proposer un modèle de traitement de ce type de polysémie lexicale. Ce modèle s'appuie, pour les aspects prédicatifs (sous-catégorisation verbale), sur une hiérarchie des différents types d'arguments, et pour la polysémie nominale, sur une formalisation de la notion de type pointé introduite par le lexique génératif [29].

Ce modèle permet un traitement robuste et réutilisable de problèmes difficilement traitables par une approche éclatée de la polysémie nominale, même au sein d'un même système de dialogue. Dans [50], nous avons montré par exemple que les noms d'action ambigus entre action et objet résultat d'une action ("acquisition") nécessite un traitement tel que celui proposé pour la polysémie logique.

6.5 Aspects génération

Notre travail en commun avec Laurence Danlos sur la génération en langue naturelle a essentiellement porté sur l'aspect planification de la structures des documents. Suite aux travaux de thèse de Laurent Rousarie [Rou00], nous avons adopté la SDRT [Ash93] (que nous utilisons déjà pour l'analyse de la co-référence événementielle) comme format de sortie. Nous avons montré dans un premier temps que ce format de sortie était adapté aux modules de génération proprement dits ; dans un second temps, nous avons décrit un algorithme qui passe d'une représentation factuelle (décrite dans un sous ensemble de la logique du premier ordre) à des ensembles de SDRS c'est à dire à des représentations en sémantique dynamique du discours [42]. Cet algorithme engendre autant de solutions que possible afin de ne pas être bloqué par l'inévitable incomplétude du générateur. Pour illustrer ce point, à partir d'une représentation

-
- [Pus95] J. PUSTEJOVSKY, *The generative Lexicon*, The MIT Press, 1995.
- [GJ96] D. GODARD, J. JAYEZ, « Types Nominaux et Anaphores ; le cas des objets et des événements », *in: Anaphores temporelles et (in-)cohérence*, Cahiers Chronos, 1996.
- [Kle99] G. KLEIBER, *Problèmes de sémantique : la polysémie en questions, sens et structures*, Villeneuve d'Ascq : Presses Universitaires du Septentrion, 1999.
- [Cru95] D. CRUSE, « Polysemy and related phenomena from a cognitive linguistic viewpoint », *in: Saint-Dizier, P. and Viegas, E. (eds.)*, 1995.
- [AP00] N. ASHER, J. PUSTEJOVSKY, « The Metaphysics of Words in Context », submitted to *Journal of Logic, Language and Information*, 2000, <http://www.dla.utexas.edu/depts/philosophy/faculty/asher/papers/MWC.ps>.
- [PK00] M. PINKAL, M. KOHLHASE, « Feature Logic for Dotted Types: A Formalism for Complex Words Meanings », *ACL*, 2000.
- [Rou00] L. ROUSSARIE, *Un modèle théorique d'inférences de structures sémantiques et discursives dans le cadre de la génération automatique de textes*, Thèse de doctorat en linguistique, Université Denis Diderot, Paris 7, 2000.
- [Ash93] N. ASHER, *Reference to Abstract Objects in Discourse*, Kluwer, Dordrecht, 1993.

paraphrasable par “Marie a fait une crise de nerf à cause du départ de Luc”, l’algorithme engendre des structures discursives paraphrasables par : “Le départ de Luc a provoqué une crise de nerf de Marie”, “Luc est parti, de ce fait, Marie a fait une crise de nerf”, etc.

6.6 Dialogue multimodal

Multimodalité et pertinence Dans le dialogue spontané, le nombre d’expressions langagières possibles pour référer à un objet est très élevé. Lorsque la communication peut se faire à l’aide des deux modalités que sont la parole et le geste de désignation, le nombre d’expressions multimodales possibles est encore plus élevé. En dialogue homme-machine, ce nombre explose si l’on tient compte de la complexité du support de la communication qu’est la scène affichée à l’écran. La perception visuelle qu’a l’utilisateur de cette scène peut en effet l’amener à utiliser des expressions plus ambiguës, qui seront interprétées correctement par intervention de la saillance visuelle et des groupements perceptifs d’objets [Wer23]. Nous proposons dans [52] une synthèse de ces notions de saillance et de groupements perceptifs, ainsi qu’un premier pas dans leur intégration dans un modèle général d’interprétation des expressions référentielles multimodales, en particulier dans le modèle des domaines de référence de l’équipe.

Face à cette variabilité se fait ressentir le besoin de disposer d’heuristiques permettant de réduire l’espace de recherche, aussi bien en compréhension qu’en génération. La théorie de la pertinence [SW95] nous semble pouvoir fournir un tel critère. Disposer de la capacité d’évaluer la pertinence d’une expression permettrait à un système de dialogue non seulement — en génération — de faire des choix efficaces parmi les modalités et les expressions disponibles, mais également — en compréhension — de remettre en question les résultats d’un calcul de référents en cas de faible pertinence, ainsi que de repérer les comportements inattendus ou les intentions particulières de l’utilisateur. Nous proposons dans un chapitre de livre actuellement en impression [30] un ensemble de règles qui permettra d’aboutir à une évaluation de la pertinence généralisable en contexte multimodal.

Intégration sémantique des modalités Avant de faire intervenir la pertinence, il faut tout d’abord disposer de méthodes formelles pour résoudre les références aux objets, problème qui peut s’avérer très complexe en dialogue multimodal spontané. Les approches classiques tendent soit à considérer que le geste de désignation donne directement les référents, soit à procéder à la fusion des informations provenant des différents média. Nous allons plus loin en supposant que la fusion se fait après une certaine compréhension, donc à un niveau sémantique. Nous avons ainsi proposé dans une conférence nationale [52] un mécanisme de compréhension multimodale. Ce mécanisme consiste à confronter les formes logiques du geste et de la parole, obtenues l’une et l’autre par une analyse syntaxique puis sémantique des informations provenant de la modalité considérée. Les travaux effectués en analyse du geste dans l’équipe nous ont permis d’aboutir à une formalisation logique du geste dans une application et un paradigme d’interaction donnés. Les travaux classiques en sémantique formelle ainsi que ceux en cours dans l’équipe nous permettent de disposer d’une forme logique des énoncés oraux. La fusion proprement dite

[Wer23] M. WERTHEIMER, *Untersuchungen zur Lehre von der Gestalt II*, 1923, ch. Psychologische Forschung.

[SW95] D. SPERBER, D. WILSON, *Relevance: Communication and Cognition*, (2nd edition) Blackwell, Oxford, 1995.

permet d'identifier des hypothèses de correspondance entre mots et geste, puis des hypothèses de référents. Ces hypothèses sont ensuite intégrées à la forme propositionnelle de l'énoncé multimodal complet, exprimée puis traitée par le système de dialogue pour répondre aux attentes de l'utilisateur.

Ce principe consistant à effectuer une fusion à un niveau sémantique n'est pas très répandu dans la communauté internationale travaillant sur les interfaces multimodales. Nous avons l'intention de le défendre dans le projet MIAMM (cf. plus loin) qui fournira ainsi un cadre de validation de notre approche. Ce travail devrait également conduire à la définition d'un format unifié de représentation de données multimodales, MMIL (Multi-Modal Interface Language) dans le projet MIAMM.

6.7 Inférence dans le traitement automatique des langues

L'inférence est une composante essentielle du système linguistique : elle contribue à résoudre l'ambiguïté (lexicale, de rattachement, de projection de présuppositions etc.), à déterminer la valeur des éléments anaphoriques (pronoms, GN défini, noms composés etc) et à lever la sous-spécification (e.g. de portée des quantificateurs).

Pendant longtemps pourtant, les techniques de raisonnement automatique étaient trop inefficaces, pour pouvoir supporter autre chose que des systèmes prototypiques. Au cours de la dernière décennie, cette situation s'est radicalement modifiée et les techniques visant à représenter et à utiliser les connaissances encyclopédiques ont atteint un niveau de performance qui permet d'envisager leur utilisation pour la langue naturelle. A partir de ces observations, nous poursuivons les objectifs suivants :

- Identification des problèmes linguistiques qui peuvent être résolus par le biais de techniques de raisonnement automatique.
[27] montre en particulier, comment les critères de consistance, d'informativité et de minimalité peuvent être formalisés et vérifiés par le biais de l'utilisation des prouveurs de théorèmes.
- Optimisation des techniques de raisonnement automatique appropriées.
- Recherche théorique sur les formalismes de représentation et les modèles d'inférence. Nos investigations se concentrent sur la logique modale [13] et, en particulier, sur la logique hybride, une forme de logique modale dans laquelle il est possible de faire référence aux mondes, au temps ou aux états ^[Bla00b] [19]. Nous étudions la logique hybride dans une perspective de complexité théorique ^[ABM00], de théorie de preuve ^[BT99, Bla00a] [21] et, plus généralement, mathématique [17]. Des implémentations expérimentales ont été réalisées

[Bla00b] P. BLACKBURN, « Representation, Reasoning, and Relational Structures: a Hybrid Logic Manifesto », *Logic Journal of the IGPL* 8, 2000, p. 339–365.

[ABM00] C. ARECES, P. BLACKBURN, M. MARX, « The Computational Complexity of Hybrid Temporal Logic », *Logic Journal of the IGPL* 8, 5, 2000, p. 653–679.

[BT99] P. BLACKBURN, M. TZAKOVA, « Hybrid languages and temporal logic », *Logic Journal of the IGPL* 7, 1, 1999, p. 27–54.

[Bla00a] P. BLACKBURN, « Internalizing labelled deduction », *Journal of Logic and Computation* 10, 2000, p. 136–168.

[36]. Par ailleurs, des liens avec le langage naturel ont été explorés [23, 20, 37, 25]. Plus récemment, nous avons étudié la logique hybride du premier ordre [18, 38] qui sera notre centre d'intérêt à court terme, car, dans une perspective de traitement du langage naturel, elle nous paraît vitale.

- Spécification et implémentation d'une architecture pour le traitement automatique des langues intégrant inférence, analyse (production à partir d'un texte du sens de ce texte [?]) et génération (production à partir d'un but communicatif d'un texte réalisant ce but [GS01]).

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Thomson-LCR

Thomson-LCR a accueilli plusieurs personnels (thèses CIFRE, post-doc industriels et embauches) issus de l'équipe. Les thèmes de collaboration ont porté sur le dialogue multimodal et, plus particulièrement sur l'interprétation contextuelle d'expressions référentielles. La dernière action en date (post-doc d'Olivier Grisvard, janv-dec 2000) a vu la réalisation en commun, en parallèle avec les travaux de thèse de Susanne Salmon-Alt, d'un module de calcul référentiel reposant sur les domaines de référence.

7.2 EADS/MSI

Notre collaboration récente (déc. 2000) avec EADS/MSI (Matra Systèmes d'Information), vise à expérimenter les méthodes d'analyse locales développées dans l'équipe dans le cadre d'applications de recherche d'information dans des documents spécialisés (sécurité informatique, dossiers médicaux). Elle est l'occasion pour nous de mettre l'accent sur la paramétrisation de modèles linguistiques (thèse de Benoit Crabbé et de Dalila Beckouche). Initiée dans le cadre d'un projet impliquant la DGA (projet Vulcain), elle se poursuit par un projet RNTL récemment labélisée.

7.3 Cinergie

Cinergie (<http://www.loria.fr/projets/CINERGIE>) est un projet industriel contracté dans le cadre du programme PRIAMM (programme pour la recherche et l'innovation dans l'audiovisuel et le multimédia, www.cnc.fr/priamm), financé par le Centre National de le Cinématographie et le secrétariat d'état à l'industrie. Ce projet a débuté le 1 septembre 2001 et est d'une durée de 18 mois. Il est mené en partenariat avec le CERIME (Centre d'Etudes et de Recherches Interdisciplinaires sur les Médias en Europe, Université Robert Schuman, Strasbourg) et la société Lucid-IT. LED en est le partenaire coordinateur.

Comme son nom le suggère, le projet CINergie vise au développement d'une application de type collecticiel (ou groupware). Il s'agit de mettre en place un environnement de gestion

[?] *** ERROR: citation 'BlaBos' undefined ***

[GS01] C. GARDENT, K. STRIEGNITZ, « Generating Indirect Anaphora », *in: Proceedings of the Fourth International Workshop on Computational Semantics (IWCS-4)*, 2001.

collaborative de bases de données cinématographiques européennes, à la fois multilingues et multimédias. L'équipe LED apporte son expertise dans les domaines :

- de la standardisation XML de documents électroniques, notamment concernant la norme TMF (ISO 16642)
- des outils de traitement collégial et de diffusion de l'information
- de la gestion du multilinguisme (projets MLIS/DHYDRO, HLT/SALT)

Un ingénieur expert (Annalisa Ricci) a été recruté spécifiquement sur ce projet.

7.4 Lucid-IT

Au titre des actions industrielles, nous pouvons aussi signaler la création de l'entreprise de technologie Lucid-IT par d'anciens membres de l'équipe. Les objectifs de Luci-IT permettent notamment de valoriser les travaux menés dans l'équipe depuis plusieurs années dans le domaine de la gestion de ressources linguistiques (textes et terminologies multilingues notamment).

7.5 Actions régionales

Diatelic Le projet Diatelic est une action transversale du Loria faisant intervenir les équipes Maia, Trio et Langue et Dialogue. Lancée initialement (1996) sous l'égide du programme «habitat intelligent» du département SPI du CNRS, ce projet associe des médecins de l'Altir (Association Lorraine de Traitement de l'Insuffisance Rénale) ainsi qu'un médecin conseil. Il vise à mettre en œuvre une plateforme de télésurveillance de malades dialysés à domicile permettant de contrôler quotidiennement leur état général et de déclencher si besoin est des alertes préventives de possibles pathologies. En 1999, bénéficiant du soutien du pôle européen de santé, une expérimentation médicale a été lancée pour comparer une population de patients équipés du système que nous avons implanté et une population de référence.

Projet « Ingénierie des langues, du document et de l'Information Scientifique, Technique et Culturelle » dans le cadre du plan État-Région Lorraine Les objectifs que nous nous fixons dans le cadre de ce projet centré sur l'ingénierie des langues, du document et de l'information scientifique, technique et culturelle sont :

- Conforter et créer des synergies entre la recherche (LORIA, INIST, ATILF (ex-INALF), et laboratoires universitaires), les offreurs de contenus (Universités, BRGMS), les développeurs de solutions (INIST, Berger-LevraultS), les formations initiales (universités de Nancy et de Metz) et les utilisateurs : le monde économique et celui de la formation professionnelle ;
- Structurer au niveau régional un centre de compétence, pour proposer et diffuser des solutions techniques innovantes et favoriser la valorisation des résultats de recherche ;
- Conforter des activités économiques en ce domaine en donnant des atouts supplémentaires aux industries lorraines dans le cadre de la société de l'information.

Notre objectif concret durant les trois prochaines années (2001-2003) est de mettre en place une double plateforme informatique de gestion, distribution et accès à des informations textuelles :

- l'une orientée vers l'Information Scientifique et Technique,
- l'autre orientée vers la gestion de ressources linguistiques (textes annotés, lexiques, dictionnaires, outils d'ingénierie linguistique).

Une partie des outils et des logiciels de base serait, bien entendu, commune à ces deux plateformes. Pour permettre une intégration aisée des divers composants logiciels, il convient de s'appuyer sur des formats d'échange entre les divers composants fondés sur des normes clairement définies (XML, TEI, etc.) et d'intégrer au mieux les compétences des diverses équipes participantes.

LORiatech Par l'intermédiaire de la structure régionale LORiatech, cette année, nous avons collaboré avec la Société ALTACOM (Eulmont, 54) sous la forme de conseils autour des technologies JAVA/XML.

ANVAR Par l'intermédiaire de L'ANVAR, cette année, nous avons collaboré avec la Société Kika Medical Multimédia (Nancy) sous la forme d'encadrement de personnel dans le cadre d'une bourse CORTECH.

7.6 Actions nationales

Collaboration Loria-ATILF En marge du projet Silfide, nous avons entamé depuis le début de l'année, une collaboration plus étroite avec l'équipe « nouveaux corpus » de l'ATILF. L'objectif de cette collaboration est d'accompagner le nettoyage et la normalisation d'un ensemble de textes (théâtre classique, poésie) en conformité avec les choix (XML et TEI) définis au sein de Silfide. Nous avons ainsi abouti à une proposition d'en-tête simplifiée permettant de documenter de manière uniforme l'ensemble des textes choisis. A partir de cette expérience, l'objectif est maintenant d'affiner progressivement les spécifications de la chaîne d'acquisition de textes nouveaux à l'ATILF pour produire le plus en amont possible des documents normalisés.

ARC RLT L'Action de Recherche Concertée «Ressources Linguistiques pour les TAG» (ARC RLT) est une collaboration entre le LORIA (LED, Calligramme), l'INRIA Rocquencourt (ATOLL) et l'Université Paris 7 (TALANA). Elle a pour objectif d'examiner les problèmes d'acquisition et de représentation des ressources dans le cadre des grammaires d'arbres adjoints (TAG – Tree Adjoining Grammars) dans un contexte de grammaires à large couverture linguistique et des lexiques associés à ces grammaires.

7.7 Actions internationales

HLT/Salt Le projet HLT/Salt (<http://www.loria.fr/projets/SALT>), sous la coordination de Gerhard Budin de l'Université de Vienne, vise à définir un ensemble de modules génériques ('middleware') pour la manipulation et la transformation de terminologies et de lexiques multilingues. Il repose sur la définition d'un format unique XLT qui intègre à la fois les possibilités de représentation de formats strictement terminologiques tels que Martif (ISO 12200) et les

données nécessaires à l'échange de lexiques informatisés pour des applications telle que la traduction automatique. De fait, ce projet travaille en liaison étroite avec le consortium OLIF (<http://www.olif.net>) qui regroupe l'essentiel des industriels du domaine. Forte de l'expérience du projet MLIS/Dhydro, l'équipe Langue et Dialogue assure la coordination technique de ce projet, en lien avec la définition de la nouvelle norme ISO 16642 dont nous avons la responsabilité. Nous définissons en particulier des méthodes permettant de générer automatiquement d'une part des filtres de rétroconversion de bases terminologiques et lexicales existantes et d'autre part des schémas (au sens XML) de vérification des documents ainsi générés.

IST/MIAMM Le projet européen MIAMM (Multidimensional Information Access using Multiple Modalities), dont nous assurons la coordination, vise à montrer la faisabilité d'un dialogue homme-machine intégrant un dispositif à retour d'effort (haptique). Ce projet, qui a débuté le 1er septembre 2001 et se finira en février 2004, est l'occasion pour notre équipe d'intervenir plus particulièrement sur les aspects suivants :

- définition d'une architecture logicielle ouverte reposant sur un format normalisé d'échange entre modules de traitement ;
- paramétrisation lexicale d'une composante d'analyse syntaxique, en intégrant en particulier les travaux menés autour de la thèse de Benoit Crabbé et en explorant le rôle que peut jouer dans ce cadre le concept de méta-grammaire ;
- intégration d'un module de calcul de la référence qui tienne compte des informations fournies par le dispositif à retour d'effort en complément des informations linguistiques ou graphiques disponibles.

7.8 Actions de normalisation

TagML, schéma de codage XML pour grammaires TAG Lexicalisées Dans le cadre d'un groupe de travail comprenant principalement TALaNa (Université de Paris 7), l'ENST (Paris, France), l'INRIA (Rocquencourt, France), le LORIA (Nancy, France) et le DFKI (Saarbrücken, Allemagne), nous avons mené une réflexion sur la normalisation des ressources prenant part dans la conception de grammaires TAG. Cette réflexion a débouché sur la spécification d'une DTD (Document Type Definition) pour le codage de schémas, schémas d'arbres élémentaires lexicalement sous-spécifiés. Les objectifs de la recommandation TagML (TAG Markup Language) sont principalement :

- de se focaliser sur un codage normalisé des ressources construit sur la recommandation XML et reprenant en partie des directives de la TEI ;
- de permettre l'aspect incrémental des grammaires ;
- de simplifier la gestion d'une grammaire en exploitant le principe de ressources virtuelles ;
- de coder des fragments de données redondantes dans le but de partager entre les arbres des structures de traits communes, des équations de traits et des sous-arbres et ainsi d'optimiser les traitements et analyses.

A terme, une des perspectives pour TagML est (1) de fournir une plateforme pour la conception de grammaires TagML comprenant les arbres élémentaires et les lexiques et (2) sur la base du parser TAG de Patrice Lopez, de développer un interpréteur TagML.

MSC/Martif De septembre à novembre 1999, et sur la base de l'expérience acquise au sein du projet Dhydro, nous avons participé en tant qu'experts à la définition d'une proposition de norme ISO (item ISO/TC 37/SC 3 N 363) sur le transfert en aveugle de données terminologiques multilingues : « Computer Applications in Terminology – Representation format for terminological data collections – MARTIF-compatible with specified constraints ». Ce travail repose en partie sur la norme Martif (ISO 12200).

ISO 16642 Depuis août 2000, l'équipe Langue et Dialogue est responsable de la définition de la future norme ISO 16642 (Computer applications in terminology - Terminological markup framework (TMF)). TMF est un modèle générique permettant de spécifier les contraintes portant sur tout langage particulier de description de données terminologiques multilingues informatisées. Indépendant d'une réalisation particulière (sous la forme d'un type de document XML par exemple), il permet de définir les conditions d'interopérabilité entre deux langages, ainsi que les procédures à mettre en œuvre pour décrire les filtres permettant de passer de l'un à l'autre. Ce modèle vise en particulier à intégrer les travaux déjà effectués dans le cadre de la norme ISO 12200 (le format Martif peut être vu comme une application de TMF) ainsi qu'à l'Université de Rennes autour du format Geneter. Il s'articule étroitement avec la révision actuelle de la norme ISO 12620 (Computer applications in terminology - Data Categories), pour laquelle nous avons proposé une modélisation en RDF (Resource Description framework) permettant d'en formaliser l'articulation avec TMF.

7.9 Relations bilatérales internationales

NSF - CNRS Depuis début 1999 et pour trois ans un projet NSF-CNRS vient supporter les échanges réguliers que nous entretenons avec le Vassar College (Nancy Ide). Les thèmes de la collaboration portent plus particulièrement sur :

- les structures discursives et la référence,
- l'annotation de ressources linguistiques,
- la modélisation des dictionnaires électroniques.

France - Vietnam La thèse de Ti Min Huyen, soutenue par l'A.U.P. (Agence Universitaire de la Francophonie), est l'occasion pour nous d'entamer une collaboration avec l'université des sciences naturelles de Hanoï sur la définition d'outils d'ingénierie linguistique communs pour le Français et le Vietnamien.

France - Brésil En vue d'analyse de corpus anglais, portugais et français et de développement d'un système de compréhension automatique des descriptions définies démonstratives, nous collaborons avec l'Université UNISINOS São Leopoldo (Brésil) dans le cadre du projet «CommonRefs». Ce projet est co-financé par l'INRIA et le CNPq-ProtémCC pour une durée de 2 ans à partir de septembre 2001.

Université de Sarrebruck(Allemagne)/Projet InDiGen Des échanges réguliers ont lieu avec le Département de Linguistique Informatique de l'Université de Sarrebruck (Allemagne) dans le cadre de la cotutelle de la thèse franco-allemande de Kristina Streignitz (Directeurs de thèse : Manfred Pinkal (Sarrebruck), Claire Gardent (Nancy) et Jean-Marie Pierrel (Nancy)); ainsi que dans le cadre du projet InDiGen (Integrierte Diskursgenerierung), projet financé par le DFG (Deutsche Forschungsgemeinschaft) mais dont le responsable scientifique (Claire Gardent) est à Nancy.

Université d'Amsterdam (Pays-Bas) Un financement de l'INRIA, pour une durée de 3 ans, en vue de la mise en place d'une équipe associée avec le groupe «Computational and Applied Logic Group» (CALG) de l'Université d'Amsterdam a été obtenue cette année. CALG est le plus grand des 5 groupes qui constituent l'ILLC de l'Université d'Amsterdam. L'objectif de cette association est de mettre en commun l'expertise de LED et de CALG dans le domaine des logiques pour la représentation des connaissances, les méthodes d'inférence et l'informatique linguistique en vue d'une application aux problèmes de traitement de langues naturels, avec un intérêt particulier pour l'utilisation des méthodes tabulaires (<http://www.loria.fr/projets/ledcalg/>).

Université de Taiwan Une convention de collaboration avec la National Chi Nan University (Taiwan) a été récemment mise en place. Cette collaboration porte sur le développement d'environnements de gestion de «musée numérique» fondé sur la technologie XML.

8 Diffusion de résultats

8.1 Animation de la communauté scientifique

Le projet Langue et Dialogue est fortement impliqué dans l'animation de la communauté scientifique dans les domaines qui sont les siens. Cela se traduit en particulier par :

- Co-direction de l'UMR ATILF/Université Nancy 2 (J-M. Pierrel)
- Président du comité TC37/SC4 de l'ISO (L. Romary)
- Responsabilité du projet « Ingénierie des langues, du document et de l'information scientifique, technique et culturelle » dans le cadre du CPER Lorrain (J-M. Pierrel)
- Membre du conseil scientifique de la TEI (L. Romary)
- Membre du bureau de l'EACL (European Chapter of the Association for Computational Linguistics) (C. Gardent)
- Membre du conseil scientifique de l'ACI « Travail » du MENRT (J-M. Pierrel)
- Membre nommé au groupe de réflexion sur le traitement informatique du langage auprès du Conseil Consultatif de la Langue Française (J-M. Pierrel)
- Membre du conseil de l'Ecole doctorale de Linguistique de l'Université de Paris 7 (J-M. Pierrel)
- Membre du Conseil Scientifique du Pôle de Recherche Scientifique et Technologique (PRST) Intelligence Logicielle (J-M. Pierrel)
- Membre du conseil scientifique de l'ACI GRID « Grille de ressources informatiques et de données » (J-M. Pierrel)

- Membre élu du Conseil scientifique de l’UHP, Université Nancy 1 (J-M. Pierrel)
- Membre élu du Conseil des Etudes et de la Vie Universitaire de l’Université Nancy 2 (S. Cruzlara)
- Membre élu du Conseil d’administration de l’Université Nancy 2 (A. Roussanaly)
- Membre des Commissions de spécialistes des 3 Universités de Nancy (J-M. Pierrel, A. Roussanaly, S. Cruz-Lara, C. Fay-Varnier)
- Responsabilité Scientifique du projet Franco-Allemand InDiGen (Generierung von koherenten Diskursen durch Integration von Satz und Diskursplanung) dans le cadre du Schwerpunktsprogramm “Sprachproduktion : Informationsvermittlung durch natuerlicher Sprache”. En cooperation avec le Département de linguistique informatique, Université de la Sarre, Sarrebruck (Allemagne)(C. Gardent)
- Responsabilité Scientifique du Projet “Inference et Generation automatique de la Langue Naturelle”. Region Lorraine (C. Gardent)
- Responsabilité de la formation pour le CNRS et L’INRIA au Loria (L. Romary)
- Editeur en chef du Journal of Logic, Language, and Information (P. Blackburn)
- Editeur de l’édition spéciale Hybrid Logic du Journal of Logic and Computation(P. Blackburn)
- Éditeur du bulletin du chapitre européen de l’Association pour la Linguistique Informatique (*EACL Newsletter*) (C. Gardent)
- Éditrice d’une édition spéciale du WEB-SLS ("European Student Journal on Language and Speech") consacrée aux sections linguistiques d’ESSLLI 2001 (S. Salmon-Alt)
- Rédacteur en chef du journal T.A.L. (Traitement Automatique des Langues)(C. Gardent)
- Rédacteur en chef adjoint de la Revue d’Intelligence Artificielle (RIA), Editions Hermes (J.M. Pierrel)
- Rédacteur invité pour un numéro thématique T.A.L. sur le Dialogue (C. Gardent, J-M. Pierrel)
- Membre du Comité de rédaction pour : Journal of Semantics (C. Gardent), Traitement Automatique des Langues (C. Gardent, J-M. Pierrel), Computer and the Humanities (L. Romary), Information-Interaction-Intelligence (J-M. Pierrel)
- Président, SIGSEM, ACL Special Interest Group in Computational Semantics (P. Blackburn)
- Présidente de la section “Language and Computation” pour ESSLLI01 (13ème Ecole Européenne d’été pour La langue, la logique et l’Information ; Helsinki, Finlande) (C. Gardent)
- Co-chair de la session étudiante « Linguistique Computationelle » ESSLLI 2001 (S. Salmon-Alt)
- Conférence invitée, Context’01, Third International Conference on Computational Semantics, Dundee, Scotland, July 2001 (P. Blackburn)
- Conférence invitée, MOL 7 (Mathematics of Language 7), Helsinki, Finland. August 2001 (P. Blackburn)
- Conférence invitée, 3rd Tbilisi Conference on Logic, Language, and Computation, Borjomi, Georgia, September 2001 (P. Blackburn)
- Conférence invitée, 2001 Workshop on Modal Logic in AI, Joint German/Austrian conference on Artificial Intelligence, KI-2001, Vienna, Austria, September 2001 (P. Blackburn)

- Tutoriel invité Third Workshop on Inference in Computational Semantics. Sienna (Italie). Juin 2001 (C. Gardent)
- Tutoriel invité 39th Annual Meeting of the ACL (ACL'01) - joint with EACL'01, Toulouse (France), Juillet 2001 (C. Gardent)
- Conférence invitée Philosophy Department, University of Amsterdam (Pays Bas), Octobre 2001 (C. Gardent)
- Conférence invitée Grenoble Xerox Research Centre, Grenoble (France), Novembre 2001 (C. Gardent)
- Conférence invitée Journées RNRT, Toulouse, Mars 2001 (J-M. Pierrel)
- Conférence invitée Journée d'Étude sur le Dialogue Correctif, 03 Octobre 2001, IIRIT, Université Paul Sabatier, Toulouse (J-M. Pierrel)

Par ailleurs, plusieurs membres de l'équipe ont été membres de comités de programme des colloques dont :

- ICoS-3 (3rd International Workshop on Computational Semantics), 18-20 July, 2001, Siena, Italy
- TALN 01 (Traitement automatique du Language Naturel) 2001
- LACL 01 "4th International Conference on Logical Aspects of Computational Linguistic", Le Croisic, France
- ACL/EACL Workshop on Sharing Tools and Resources for Research and Education, Toulouse, France ;
- PorTAL (Portugal for Natural Language Processing), FARO Algarve, June 2002
- RECITAL 01 (Rencontre des Etudiants Chercheurs en Informatique pour le Traitement des Langues) 2001

8.2 Enseignement universitaire

La forte composante universitaire de l'équipe nous conduit à être très impliqués dans les structures d'enseignement universitaire. On peut noter en particulier :

- co-responsabilité de la filière « Perception, raisonnement et Traitement automatique des langues » du DEA d'Informatique de Lorraine (J-M. Pierrel)
- responsabilité du DEA « Modélisation et simulation des espaces bâtis » (J-M. Pierrel)
- responsabilité de la Licence et Maîtrise «Sciences Cognitives» à l'université Nancy 2 (D. Coulon)
- direction des études adjointe à l'Ecole Nationale de Géologie (C. Fay-Varnier)
- participation à l'option Bio-informatique dans la maîtrise « Biologie Cellulaire et Physiologie » de l'université Henri Poincaré (M. Smail et M-D. Devignes)
- participation à la filière PRTAL du DEA d'informatique (B. Gaiffe, L. Romary)
- participation à la licence et Maîtrise «Sciences Cognitives» à l'université Nancy 2 (D. Coulon, A.Roussanaly, S. Salmon-Alt, E. Jacquy, H. Manuélian)
- participation au DESS «TEXTE» à l'université Nancy 2 (J-M. Pierrel, A.Roussanaly, L. Romary, D. Seddah)
- participation au DESS «Industries de la langue» à l'université de Metz (S. Cruz-Lara)

8.3 Autres enseignements

Parmi les autres enseignements assurés par des membres du projet Langue et Dialogue, il convient de noter :

- Ecole d'été EACL, Computational Semantics, ESSLLI XXIII, Helsinki, Finland, 13-24 August 2001 (P.Blackburn)

9 Bibliographie

Ouvrages et articles de référence de l'équipe

- [1] P. BONHOMME, P. LOPEZ, « TagML : XML Encoding of Resources for Lexicalized Tree Adjoining Grammars », *in : Second International Conference on Linguistic Ressources and Evaluation (LREC'2000), Athens, May 2000.*
- [2] L. DANLOS, « Discours causal et rôles thématiques », *in : Des structures linguistiques à leur interprétation, mélanges en l'honneur d'André Borillo*, Rodopi, 1999.
- [3] L. DANLOS, « Event Coreference Between Two Sentences », *in : Proceedings of International Workshop on Computational Semantics, Tilburg, Pays-Bas*, 1999.
- [4] O. GRISVARD, *Modélisation et gestion du dialogue oral homme-machine de commande*, thèse de doctorat, Thèse de l'Université Henri Poincaré Nancy 1, Janvier 2000.
- [5] P. LOPEZ, *Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisées d'arbres*, thèse de doctorat, Université Henri Poincaré, Nancy 1, 1999.
- [6] J.-M. PIERREL, *Ingénierie des Langues, Traité IC2 (Information, communication et commande*, Hermes, octobre 2000.
- [7] L. ROMARY, P. BONHOMME, F. BRUNESSEAU, J.-M. PIERREL, « Silfide : A System for Open Access and Distributed Delivery of TEI Encoded Documents », *Computers and the Humanities* 33, 1-2, avril 1999, p. 31–38.
- [8] L. ROMARY, P. BONHOMME, « Parallel Alignment of Structured Documents », Kluwer Academic Press, 2000.
- [9] L. ROMARY, *Langue et dialogues - perspectives de recherche*, Habilitation à diriger des recherches, décembre 1999.
- [10] G. SABAH, J. VIVIER, A. VILNAT, J.-M. PIERREL, L. ROMARY, *Machine, langage et dialogue*, L'Harmattan, 1998.
- [11] A. SCHAAFF, *Conception et développement d'une application de recherche d'informations génomiques sur Internet. Introduction de la multimodalité.*, Thèse cnam, Metz, juin 2000.

Livres et monographies

- [12] P. BLACKBURN, M. DE RIJKE, Y. VENEMA, *Modal Logic, Cambridge Tracts in Theoretical Computer Science, 53*, Cambridge University Press, juin 2001.
- [13] P. BLACKBURN, M. DE RIJKE, Y. VENEMA, *Modal Logic*, Cambridge University Press, 2001.
- [14] P. BLACKBURN, M. KOHLHASE, *Inference in Computational Semantics*, juin 2001, ICoS-3 Workshop Proceedings, June 18-19, Siena, Italy.

Thèses et habilitations à diriger des recherches

- [15] S. SALMON-ALT, *Référence et dialogue finalisé : de la linguistique à un modèle opérationnel*, Thèse d'université, Nancy, mai 2001.

Articles et chapitres de livre

- [16] C. ARECES, P. BLACKBURN, M. MARX, « Hybrid Logic : Characterization, Interpolation and Complexity », *Journal of Symbolic Logic* 66, 3, septembre 2001, p. 977–1010.
- [17] C. ARECES, P. BLACKBURN, M. MARX, « Hybrid logics : Characterization, interpolation and complexity », *Journal of Symbolic Logic* 66(3), 2001, p. 977–1010.
- [18] C. ARECES, P. BLACKBURN, M. MARX, « Repairing the Interpolation Theorem in Quantified Modal Logic », *Annals of Pure and Applied Logic*, To appear.
- [19] C. ARECES, P. BLACKBURN, « Bringing them all together », *Journal of Logic and Computation* 11, 5, 2001, p. 657–669.
- [20] P. BLACKBURN, M. MARX, « Remarks on Gregory's "Actually" Operator », *Journal of Philosophical Logic*, To appear.
- [21] P. BLACKBURN, « Discipline as Logic : Treating Labels as First-Class Citizens », in : *Labeled Deduction*, D. Basin, D. Gabbay, S. Matthews, et L. Vigano (éditeurs), Kluwer Academic Publishers, 2000, p. 81–105.
- [22] P. BLACKBURN, « Fine Grained Theories of Time », in : *Essays on Non-Classical Logic*, H. Wansing (éditeur), *Advances in Logic*, 1, World Scientific Publishing, octobre 2001, ch. 1, p. 1–36.
- [23] P. BLACKBURN, « Fine Grained Theories of Time », in : *Essays on Non-Classical Logic*, H. Wansing (éditeur), World Scientific, 2001, p. 1–36.
- [24] P. BLACKBURN, « Modal Logic as Dialogical Logic », *Synthese* 127, 1/2, avril 2001, p. 57–93.
- [25] P. BLACKBURN, « Modal Logic as Dialogical Logic », *Synthese* 127, 2001, p. 57–93.
- [26] D. DUCHIER, C. GARDENT, « Tree Descriptions, Constraints and Incrementality », in : *Computing Meaning*, R. M. H. Bunt et E. Thijsse (éditeurs), *Studies in Linguistics and Philosophy Series*, 2, Kluwer Academic Publishers, 2001.
- [27] C. GARDENT, B. WEBBER, « Towards the Use of Automated Reasoning in Discourse Disambiguation », *Journal of Logic, Language and Information*, 2001.
- [28] N. IDE, L. ROMARY, « Encoding syntactic annotation », in : *Treebanks*, A. Abeillé (éditeur), Kluwer Academic Publisher, 2001.
- [29] E. JACQUEY, « A unified approach to lexical ambiguity of dot objects », *Natural language generation, analysis and learning*, 2001.
- [30] F. LANDRAGIN, A. DE ANGELI, F. WOLFF, P. LOPEZ, L. ROMARY, « Relevance and Perceptual Constraints in Multimodal Referring Actions », in : *Information sharing*, K. van Deemter et R. Kibble (éditeurs), CSLI Publications, 2001.
- [31] L. ROMARY, « Un modèle abstrait pour la représentation de terminologies multilingues informatisées », *Cahiers de GUTenberg*, 39-40, mai 2001, p. 81–88.
- [32] S. SALMON-ALT, « Du corpus à la théorie : l'annotation (co-)référentielle. Traitement Automatique des Langues », *Revue T.A.L., Hermès, Paris*, 2, vol.42, 2001.
- [33] A. TROGNON, D. COULON, « La modélisation des raisonnements générés dans les interlocutions », *Langages*, 2001.

Communications à des congrès, colloques, etc.

- [34] P. BERNARD, C. BERNET, J. DENDIEN, J.-M. PIERREL, G. SOUVAY, Z. TUCSNAK, « Ressources linguistiques informatisées de LATILF », *in* : *TALN 2001, Tours*, juillet 2001.
- [35] P. BERNARD, J. DENDIEN, J. LECOMTE, J.-M. PIERREL, G. SOUVAY, Z. TUCSNAK, « Les ressources informatisées de LATILF pour l'étude du français : dictionnaires, encyclopédie, bases textuelles et logiciels d'exploitation », *in* : *séminaire Corpus et ressources de ILLF, Paris*, 2001.
- [36] P. BLACKBURN, A. BURCHHARDT, S. WALTER, « Hydra : a tableau-based prover for basic hybrid logic (System Description) », *in* : *Proceedings of M4M-2 : Second International Workshop on Methods for Modalities*, M. de Rijke, C. Areces (éditeurs), ILLC, University of Amsterdam, 2001.
- [37] P. BLACKBURN, M. MARX, « Quantified Hybrid Logic and Natural Language », *in* : *Proceedings of the Thirteenth Amsterdam Colloquium*, R. van Rooy, P. Dekker (éditeurs), ILLC, University of Amsterdam, 2001.
- [38] P. BLACKBURN, M. MARX, « A Tableau System for Quantified Hybrid Logic », *in* : *Proceedings of M4M-2 : Second International Workshop on Methods for Modalities*, M. de Rijke, C. Areces (éditeurs), ILLC, University of Amsterdam, 2001.
- [39] C. COLIN, S. CRUZ-LARA, A. DURAND-SALMON, C. MATHULIK, V. VIGNERON, « Mise en application de la technologie XML au sein du projet "TELA" », *in* : *VSST'2001 (Veille Stratégique, Scientifique et Technologique), Barcelone, Espagne*, octobre 2001.
- [40] S. CRUZ-LARA, J.-L. HUSSON, A. SCHAAFF, « XML-based applications for accessing and sharing natural language resources », *in* : *2nd Sino-Franco Workshop on multimedia and XML*, mars 2001.
- [41] S. CRUZ-LARA, « Ressources linguistiques textuelles : l'apport de nouvelles technologies de l'information et de la communication », *in* : *Colloque National de la Recherche Universitaire dans les I.U.T.*, juin 2001.
- [42] L. DANLOS, B. GAIFFE, R. LAURENT, « Document Structuring à la SDRT », *in* : *8th European Workshop on Natural Language Generation, Toulouse*, juillet 2001.
- [43] G. DUFFING, « Thematico-Visual Image Retrieval : How to Deal With Partially Indexed Corpora », *in* : *Internet Imaging II, San Jose, CA*, SPIE, janvier 2001.
- [44] C. GARDENT, S. THATER, « Generating with a Grammar Based on Tree Descriptions : a Constraint-Based Approach », *in* : *39th Annual Meeting of the Association for Computational Linguistics*, juillet 2001.
- [45] M. GUPTA, L. ROMARY, « Dynamically Generated Interfaces in XML Based Architecture », *in* : *UIML Europe 2001, Paris, France*, mars 2001.
- [46] N. IDE, L. ROMARY, T. ERJAVEC, « A Common XML-based Framework for Syntactic Annotations », *in* : *1st NLP and XML Workshop, Tokyo, Japan*, novembre 2001.
- [47] N. IDE, L. ROMARY, « A Common Framework for Syntactic Annotation », *in* : *ACL 2001, Toulouse, France*, juillet 2001.
- [48] N. IDE, L. ROMARY, « Standards for Language Resources », *in* : *IRCS Workshop on Linguistic Databases, Philadelphia, USA*, Steven Bird, Peter Buneman and Mark Liberman, décembre 2001.
- [49] N. IDE, L. ROMARY, « Towards a common framework for linguistic annotation », *in* : *SLE workshop on Linguistic Interpretation of Corpora (LINC 2001), Leuven, Belgium*, août 2001.
- [50] E. JACQUEY, « Ambiguïté lexicale des verbaux d'actions et catégorie conceptuelle », *in* : *Representations du sens linguistique, Bucarest, Roumanie*, 2001.

-
- [51] F. LANDRAGIN, N. BELLALEM, L. ROMARY, « Compréhension automatique du geste et de la parole spontanés en communication homme-machine : apport de la théorie de la pertinence », *in : Oralité et Gestualité, Aix-en-Provence, France*, 2001.
- [52] F. LANDRAGIN, N. BELLALEM, L. ROMARY, « Visual Saliency and Perceptual Grouping in Multimodal Interactivity », *in : Workshop on Information Presentation and Natural Multimodal Dialogue, Verona, Italy*, 2001.
- [53] M. MARTINEZ, S. CRUZ-LARA, P. DE LA FUENTE, J.-C. DERNIAME, « La génération de documents virtuels dans les bibliothèques électroniques : application aux textes juridiques », *in : VSST'2001 (Veille Stratégique, Scientifique et Technologique), Barcelone, Espagne*, octobre 2001.
- [54] L. ROMARY, M. VAN CAMPENHOUDT, « Normalisation des échanges de données en terminologie : les cas des relations dites conceptuelles », *in : TIA (Terminologie et Intelligence Artificielle), Nancy, France*, mai 2001.
- [55] L. ROMARY, « An abstract model for the representation of multilingual terminological data : TMF - Terminological Markup Framework », *in : TAMA 2001, Anvers, Belgium*, février 2001.
- [56] S. SALMON-ALT, « Interpreting "other" : from Cognitive Grammar to multimodal dialogues », *in : SemPro - Workshop on cognitively plausible models of semantic processing, Edinburgh, UK*, juillet 2001.
- [57] S. SALMON-ALT, « Reference Resolution within the Framework of Cognitive Grammar », *in : International Colloquium on Cognitive Science, San Sebastian, Spain*, mai 2001.
- [58] J.-P. THOMESSE, D. BELLOT, A. BOYER, E. CAMPO, M. CHAN, F. CHARPILLET, J. FAYN, C. LESCHI, N. NOURY, V. RIALLE, L. ROMARY, P. RUBEL, N. SELMAOUI, F. STEENKESTE, G. VIRONE, « Integrated Information Technologies for patients remote follow-up and homecare », *in : HealthCom 2001, l'Aquila, Italie*, juillet 2001.

Rapports de recherche et publications internes

- [59] B. CRABBÉ, « Organisation des propriétés syntaxiques du lexique verbal : une structure arborescente », *Mémoire de DEA*, Université Paris IV, septembre 2001.