Université de Bretagne Sud

Pour une ingénierie des langues plus linguistique

Réflexions à la lumière des recherches du laboratoire VALORIA

Mémoire présenté par

Jean-Yves Antoine

En vue d'obtenir le diplôme d'

Habilitation à diriger les recherches — spécialité informatique

Soutenance : le 28 novembre 2003

Jury

Jean Caelen	CLIPS-IMAG, Grenoble
Jacques Siroux	IRISA, LLI/CORDIAL, Lannion
Karine Krüger-Thielmann	SfS, Universität Tübingen
Jean-Marie Pierrel	LORIA & ATILF, UHP, Nancy
Françoise Gadet	U. Paris-X Nanterre, Paris
Franck Poirier	VALORIA, UBS, Vannes

Sommaire

Préambule :

L'ingénierie des langues au VALORIA

Si parva licet componere magnis
(S'il est permis de comparer les petites choses aux grandes)

Virgile, Les Géorgiques, II, 489

1. DE LA RECHERCHE AU SEIN DES UNIVERSITES NOUVELLES

L'habilitation à diriger les recherches est l'occasion pour tout chercheur de faire un bilan sur ses activités de recherche. Après avoir préparé mon doctorat à l'Institut de la Communication Parlé (INPG, Grenoble, 1994) et poursuivi des recherches post-doctorales au CLIPS-IMAG (U. Joseph Fourier, Grenoble), j'ai rejoint en septembre 1996 le laboratoire VALORIA de l'Université de Bretagne Sud (Vannes) en qualité de Maître de Conférences. J'y ai poursuivi depuis mes activités de recherche, dans un contexte stimulant mais assez inhabituel.

En effet, le VALORIA, laboratoire de recherche en informatique de la plus jeune université de France, ne développait pas d'activité en ingénierie des langues à mon arrivée. J'ai donc intégré une équipe travaillant sur l'Interaction Homme - machine (EQUIPAGE, dirigée alors par Franck Poirier), au sein de laquelle j'ai entrepris de créer un groupe de recherche en traitement automatique des langues (groupe CORAIL). La thématique principale du groupe, la communication langagière homme - machine, restait en cohérence avec les autres activités de l'équipe. Je dois remercier Franck Poirier pour la liberté et la confiance qu'il m'a toujours accordées dans la création de ce nouveau groupe de recherche[1].

Une part non négligeable de mes activités fut ainsi consacrée à la structuration et au développement de ce groupe de recherche dans un laboratoire qui n'était alors pas reconnu par le ministère, ne disposait pas de personnel administratif et n'était pas encore laboratoire d'accueil d'un DEA !

Sept ans plus tard, ce groupe a atteint une taille critique (trois enseignants - chercheurs, trois doctorants ou jeunes docteurs actuellement ATER ou PRAG) qui lui permet de poursuivre les objectifs qu'il s'assigne. Il commence, je l'espère, à voir son travail reconnu par la communauté scientifique. Ce développement, je le dois à plusieurs soutiens institutionnels (bourses doctorales, financement de projets) : la région Bretagne, le CNRS, le Ministère de la Recherche ainsi que l'Agence Universitaire de la Francophonie (AUF, ex-AUPELF-UREF). Ces soutiens furent essentiels pour développer nos activités et lutter contre notre relatif isolement. La pérennité du groupe CORAIL leur doit beaucoup.

La recherche au sein des jeunes universités constitue souvent une aventure difficile et prenante. Ces difficultés ne doivent cependant par faire oublier que ces universités et laboratoires en devenir constituent un espace d'opportunités pour les jeunes enseignants - chercheurs. C'est pourquoi je suis persuadé que l'Université de Bretagne Sud a constitué un cadre favorable à l'expression du programme scientifique qui me motive et que présente ce document : celui d'un dialogue homme - machine centré sur les besoins et les usages langagiers des utilisateurs.

Ce programme n'aurait jamais pu être envisagé sans le concours de toute une équipe. Qu'il me soit donc permis de dire ma reconnaissance aux collègues qui m'ont suivi dans la création d'une équipe en ingénierie des langues tout en adhérant à mon approche: Brigitte Le Pévédic,

[1] Ces remerciements s'adressent également à Jean Caelen qui m'a toujours accordé une prime à l'autonomie tout au long d'un doctorat … déjà lointain.

Sabine Letellier - Zarshenas, Jérôme Goulian, Igor Schadle, Jeanne Villaneau et Pascale Nicolas. Si, encadrement de thèses ou de projets obligent, une part non négligeable des travaux du groupe reposent sur mes propositions, la contribution personnelle de ces personnes ne saurait être ignorée.

Enfin, je ne saurais oublier les membres de ce jury d'habilitation. Je leur adresse tous mes remerciements pour m'avoir fait le plaisir et l'honneur de leur présence à cette soutenance, mais surtout pour leurs travaux — et quelquefois des conseils plus personnels — qui ont bien souvent guidé ma réflexion. La lecture de ce mémoire rendra, je l'espère, justice à leurs contributions. Elle suscitera peut-être également chez eux intérêts ou désaccords, en bref toutes ces petites choses qui nourrissent le débat scientifique et qui rendent notre métier si passionnant.

2. L' INGENIERIE DES LANGUES AU VALORIA : GROUPE CORAIL

Avant d'entrer dans une discussion scientifique reposant sur mes travaux de recherche, je vais dresser dans ce premier chapitre un bilan synthétique de mes activités depuis ma nomination en qualité de maître de conférences.

L'ingénierie des langues et la communication homme - machine sont des domaines technologiques où il n'est plus envisageable de mener une recherche en solitaire. C'est pourquoi j'ai cherché à développer au cours de ces années au VALORIA une équipe en ingénierie des langues autour de mes travaux. Pour cela, il m'a fallu :

- structurer les recherches du groupe CORAIL autour d'approches et de méthodes semblables, afin de compenser par une capitalisation des connaissances la taille relativement modeste du groupe,
- ancrer nos activités au sein de la communauté scientifique afin de rompre un isolement dû à l'absence de pôle scientifique d'importance sur Vannes,
- porter enfin notre projet scientifique au sein de cette communauté.

Mes activités de recherche sont donc intimement liées à celle du groupe CORAIL. C'est pourquoi ce chapitre concernera des activités qui ont marqué le développement de ce groupe sous ma direction, aussi bien que mon évolution personnelle depuis mon arrivée à l'Université de Bretagne Sud.

3. DEVELOPPER ET STRUCTURER LES RECHERCHES EN INGENIERIE DES LANGUES A L'UNIVERSITE DE BRETAGNE SUD

Mes recherches doctorales et post-doctorales étaient centrés sur la compréhension de parole en situation de dialogue homme - machine finalisé. Dès mon arrivée au VALORIA en 1996, j'ai cherché à généraliser les travaux du groupe CORAIL[2] à d'autres problématiques de l'interaction homme - machine, sans abandonner une approche centrée sur l'utilisateur[3]. Intégré dans un jeune laboratoire, il m'importait également que ce groupe de recherche atteigne une réelle autonomie scientifique. C'est pourquoi j'ai orienté mes travaux et ceux du groupe dans deux directions compatibles avec cette exigence d'autonomie.

1) **Nouveaux cadres applicatifs** — Nos recherches se concentrent désormais sur deux cadres applicatifs bien définis. Il s'agit tout d'abord du dialogue oral homme - machine, pour lequel je disposais d'une expérience déjà solide puisque ce domaine est en continuité logique de mes travaux de doctorat. CORAIL dispose ainsi de deux systèmes de

[2] De sa création en 1996 jusqu'en 2001, ce groupe était dénommé EQUIPAGE-LN.

[3] Ce programme scientifique était d'ailleurs partagé par l'équipe EQUIPAGE créée par Franck Poirier.

compréhension de parole pour le dialogue homme - machine finalisé qui ont été réalisés dans le cadre de deux doctorats (cf. § 3.1). A l'opposé, j'ai développé un nouvel axe de recherche centrée sur l'aide à la communication langagière pour les personnes handicapées. Cette thématique, qui relève de la communication homme - homme médiée par ordinateur[4], répond à un besoin social fort. Elle présente par ailleurs deux intérêts pour CORAIL. Il s'agit tout d'abord d'un cadre applicatif « niche » adapté à la taille d'un groupe émergent tel que le notre. Mais surtout, l'aide au handicap nécessite une prise en compte cruciale de l'utilisateur que ce soit au niveau ergonomique (interaction homme - machine) ou langagier (ingénierie des langues). Ces préoccupations recoupent complètement les motivations scientifiques de notre groupe et plus généralement de l'équipe EQUIPAGE.

2) **Linguistique de corpus et méthodologies d'évaluation** — Suivant l'approche centrée sur l'utilisateur qui me tient à cœur, il est essentiel que le développement des systèmes interactifs s'appuie sur une réflexion plus poussée à la fois en amont (analyse des usages langagiers) et en aval (évaluation détaillée des systèmes) des activités de conception. C'est pourquoi j'ai lancé un axe de recherche en linguistique de corpus portant sur le dialogue oral. Compte tenu de la disponibilité limitée des corpus oraux francophones de dialogue oral, j'ai également engagé le groupe CORAIL dans la constitution de ressources linguistiques orales répondant aux besoins de la communication homme - machine. Cette activité, qui renforce notre autonomie scientifique, s'accompagne d'une politique de libre diffusion de nos corpus. Cette démarche est restée jusqu'à présent relativement rare au sein de la communauté francophone. C'est pourquoi un des objectifs de cette politique volontariste est de faire de CORAIL un des référents en matière de ressources linguistiques orales.

   Parallèlement, j'ai souhaité faire de l'évaluation des systèmes interactifs un axe de recherche à part entière du groupe CORAIL. Outre la participation de nos systèmes à différentes campagnes d'évaluation (GDR-I3, TECHNOLANGUE), je développe ainsi une réflexion sur les pratiques de tests du domaine. La clé de cette recherche est de mettre l'évaluation des systèmes en regard avec les besoins et usages des utilisateurs. Elle a déjà donné lieu à plusieurs propositions de méthodologies d'évaluation, dont certaines en collaboration avec le CLIPS-IMAG (Jean Caelen), l'ICP (Jérôme Zeiliger) et l'IRISA (CORDIAL, Jacques Siroux).

Tout en étant clairement et volontairement délimitées, ces thématiques sont trop vastes pour être envisagées de manière isolée ! En dehors des collaborations externes sur lesquelles je reviendrai ultérieurement, mes recherches ont ainsi été menées de deux manières au cours de ces sept dernières années :

- Travaux personnels, avec l'aide éventuelle d'autres collègues du groupe (Sabine Letellier - Zarshenas, Pascale Nicolas, Jérôme Goulian) en linguistique de corpus et sur l'évaluation des systèmes interactifs,
- Encadrements de thèse (Jérôme Goulian, Igor Schadle, Jeanne Villaneau essentiellement) ou de DEA (Iadaloharivola Randria, Julien Foulon, Frédéric Lamie) sur les recherches relevant de la conception de systèmes interactifs.

[4] Contrairement à l'aide au handicap, la traduction parole-parole constitue un exemple de plus en plus étudié de communication médiée par l'ordinateur. Voir par exemple le projet NESPOLE ! financé par la communauté européenne et la *National Science Foundation* (NSF) américaine, et auquel participe le laboratoire CLIPS-IMAG : Rossato S., Blanchon H., Besacier L. (2002) Evaluation du premier démonstrateur de traduction de parole dans le cadre du projet NESPOLE ! Actes *atelier thématique « Couplage de l'écrit avec l'oral », TALN'2002*, Nancy, France. Vol 2, 149-161

Je vais tout d'abord revenir sur ces différents encadrements doctoraux.

3.1. Encadrements doctoraux et de DEA

Trois doctorats en informatique ont à ce jour été réalisés sous ma (co-)direction au sein du groupe CORAIL. Dans la continuité de mon travail de thèse, les deux premiers d'entre eux relèvent du dialogue oral homme - machine et plus précisément de la compréhension automatique de la parole.

Doctorat de Jérôme Goulian — Jérôme Goulian a soutenu son doctorat en informatique de l'Université de Bretagne Sud le 13 décembre 2002[5]. Cette thèse financée par la région Bretagne a fait suite à un *stage de DEA* sur les grammaires de dépendances[6], que j'avais co-encadré avec Damien Genthial (CLIPS-IMAG, Grenoble). L'idée originelle de la thèse provenait des convergences qui existent entre le formalisme des grammaires de dépendances et le système de compréhension ALPES que j'avais développé au cours de mon doctorat. Mon objectif était ainsi de réaliser un système qui utiliserait le formalisme des grammaires de liens et serait fondé sur l'analyse des usages langagiers en situation de dialogue oral. Au cours du doctorat, la réflexion s'est orientée vers l'intégration de techniques issues du TAL robuste. Un système de compréhension appliqué au renseignement touristique a ainsi été réalisé (ROMUS). Il repose sur une stratégie d'analyse incrémentale mixant techniques de segmentation robuste (*chunks*) et analyse de dépendances sémantiques par grammaires de liens. Cette thèse a été réalisée sous ma direction scientifique et celle de Franck Poirier comme directeur de thèse officiel. Elle a donné lieu à plusieurs publications, essentiellement dans des conférences nationales telles que TALN.

Le système ROMUS, qui présente des performances intéressantes, continue d'être développé au sein du groupe. En particulier, j'ai encadré en 2002-2003 deux stages de DEA de l'Université de Bretagne Sud (cf. infra § 3.2) qui visent à étendre et améliorer le système :

- *DEA de Iadaloharivola Randria* — Ce stage de DEA a concerné le traitement des réparations (répétitions, autocorrections) de l'oral spontané en situation de dialogue finalisé. Plus précisément, il s'agissait d'étudier l'intérêt d'un pré-traitement des énoncés reposant sur des techniques de détection des réparations par patrons (ou patterns). Jusqu'à présent, ce type de méthodes superficielles n'a pas donné lieu à une étude systématique sur le français parlé. Ce stage a permis de caractériser un ensemble limité de patrons (cf. chap. 2, § 3.3.) qui permet atteindre une excellente précision de détection en contrepartie d'un rappel moins élevé. Iadaloharivola Randria doit poursuivre ces travaux dans le cadre d'un doctorat en co-tutelle qui sera encadré par Tefy Raoelivololona (Université d'Anstiranana à Madagascar) et moi-même.
- *DEA de Julien Foulon* — Dans la perspective de l'interprétation contextuelle de la parole, ce stage de DEA concernait la résolution des anaphores pronominales. Il a étudié l'adaptation des techniques superficielles de résolution des anaphores à la problématique du dialogue oral finalisé. Jusqu'à présent, ces méthodes efficaces n'ont été utilisées que sur du texte écrit. Ce stage constituait donc une étude de faisabilité en CHM orale à partir d'analyses de corpus de dialogue pilotes. Les résultats partiels qui ont été obtenus montrent que l'adaptation directe de ces techniques n'est guère envisageable (cf. chap. 2, § 3.4).

[5] Goulian J. (2002) Stratégie d'analyse détaillée pour la compréhension automatique robuste de la parole. Thèse Université de Bretagne Sud, Vannes, France. 13 Décembre 2002. Rapport de recherche VALORIA-CORAIL-2002-03.

[6] Goulian J. (1998) Analyse Robuste du français parlé. DEA Sciences Cognitives, INPG, Grenoble, France. Juin 1998.

Toujours dans la perspective d'une analyse incrémentale du langage parlé, ces étapes de traitement (détection de réparations, résolutions des co-références anaphoriques) devraient être à terme intégrées en entrée et en sortie du système ROMUS. Julien Foulon débute ainsi un doctorat sous ma direction qui porte sur la résolution des références dans le système ROMUS. ROMUS participera ainsi, tout comme le système LOGUS décrit ci-dessous, à la prochaine campagne MEDIA d'évaluation en contexte des systèmes francophones de compréhension de parole (cf. § 5.1).

Doctorat de Jeanne Villaneau — Cette seconde thèse portant sur la compréhension de parole a été initiée dans un contexte particulier. En effet, il s'agissait originellement d'un doctorat de l'Université Rennes 1 encadré par Olivier Ridoux (IRISA/LANDES) et portant sur le lambda - calcul. Compte tenu de l'équivalence formelle qui existe entre les grammaires de dépendances et le formalisme logique des grammaires catégorielles, j'ai proposé à Olivier Ridoux et Jeanne Villaneau de transposer leurs travaux sur la problématique de la compréhension de parole appliquée au renseignement touristique. Cette réorientation s'est traduite par une inscription en doctorat en informatique de l'Université de Bretagne Sud sous la conduite d'Olivier Ridoux et moi-même. Elle a rapidement porté ses fruits sous la forme d'un système logique de compréhension — et d'interprétation — de parole (LOGUS) qui est parfaitement opérationnel. Ce système, qui constitue une adaptation réussie des approches logiques au traitement du langage parlé, répond lui aussi à une stratégie d'analyse incrémentale. Les systèmes LOGUS et ROMUS reposent ainsi sur des architectures générales comparables (segmentation en *chunks* puis analyse de dépendances globales). Ces rapprochements, qui se retrouvent dans nos recherches sur le handicap, favorisent la structuration de nos recherches. C'est ainsi que les réalisations issues du DEA d'Iadaloharivola Randria pourront être directement intégrées au système LOGUS. Pour l'heure, le doctorat de Jeanne Villaneau touche à sa fin. Elle a donné lieu à plusieurs publications, aussi bien internationales (LACL'2001) que nationales (TALN) et a été présentée publiquement en décembre 2003[7].

Le dernier doctorat réalisé au sein du groupe CORAIL concerne l'autre cadre applicatif étudié par notre équipe, à savoir l'aide à la communication langagière pour les personnes handicapées.

Doctorat d'Igor Schadle — Cadre applicatif de choix pour les recherches en robotique, la thématique de l'aide au handicap est assez peu étudiée en traitement automatique des langues naturelles (écrit). Je l'ai personnellement découverte après mon intégration au VALORIA, lors d'échanges scientifiques avec Maryvonne Abraham (ENST Bretagne, Brest) et Jean-Paul Departe (centre de rééducation fonctionnelle de Kerpape). Rapidement, il m'est apparu que le processus d'amorçage sémantique utilisé par le système ALPES pouvait être adapté à la problématique de l'aide aux handicapés aphasiques lourds (système HandiALPES). Le recrutement de Brigitte Le Pévédic, qui venait de soutenir une thèse dans ce domaine, m'a permis de faire de l'aide au handicap un axe de recherche central du groupe CORAIL. Au cours de son doctorat, Brigitte Le Pévédic avait développé un système d'aide à la saisie (HandiAS) reposant sur une modélisation probabiliste incontournable dans ce domaine[8]. Présentant des capacités de prédiction intéressantes en terme d'économie de saisie, ce système

[7] Villaneau J. (2003) Contribution au traitement syntaxico-pragmatique de la langue naturelle parlée: approche logique pour la compréhension de la parole. Doctorat l'Université de Bretagne Sud, Vannes, France. 6 décembre 2003. Rapport de recherche VALORIA-CORAIL-2003-02.

[8] Plus précisément, l'objectif de ce type de système est de fournir à la personne handicapée une liste ordonnée d'hypothèses (lettres ou mots) en fonction du contexte (texte déjà saisi). Le classement de ces prédictions ne pouvant se baser que sur des considérations numériques, on a généralement recours à des modèles probabilistes, connexionnistes ou hybrides. Ainsi, le système HandiAS reposait sur l'utilisation de transducteurs à états finis probabilistes.

a fait l'objet d'une commercialisation. De l'aveu même de sa conceptrice, il reposait cependant sur un modèle de langage relativement simple laissant place à de futures améliorations. Tout en conservant le principe d'une modélisation probabiliste, j'ai alors proposé de réaliser un système de prédiction plus fin reposant sur une analyse partielle de la structure de l'énoncé saisi. L'idée étant de baser la prédiction non plus seulement sur l'analyse de co-occurrences de mots, mais sur des structures syntaxiques minimales telle que les *chunks*. Cette proposition a fait l'objet d'un financement de thèse par la région Bretagne dont a bénéficié Igor Schadle. Elle a conduit à la réalisation d'un système d'aide à la communication appelé Sibylle. Les performances du système en terme de capacité de prédiction, de même que son utilisation par des personnes handicapés, montrent clairement que Sibylle se situe à la pointe de l'état de l'art du domaine. La portée de ce doctorat dépasse par ailleurs la simple question de l'aide au handicap. En effet, notre système rejoint les tentatives les plus récentes[9] d'amélioration des approches probabilistes en direction d'une modélisation plus profonde du langage. A ce titre, il participera à la prochaine campagne TECHNOLANGUE (projet EASy) d'évaluation des analyseurs syntaxiques du français (cf §5.2). Enfin, on relèvera une fois de plus que l'utilisation des *chunks* comme niveau élémentaire de description syntaxique est une constante de tous nos travaux. Ce doctorat est encadré par Brigitte Le Pévédic et moi-même. Le directeur de thèse officiel est Franck Poirier, qui a apporté ses compétences en matière d'interface homme - machine avec la personne handicapée. Il a donné lieu à plusieurs publications dans des revues (RIHM) et conférences nationales (JIM, TALN) et a été présenté publiquement en décembre 2003[10].

Autres encadrements doctoraux — J'interviens par ailleurs dans le cadre de deux doctorats réalisés en dehors du laboratoire VALORIA. Ces activités résultent directement de mes collaborations avec d'autres centres de recherches.

Tout d'abord, j'ai régulièrement suivi les travaux de thèse de **Mohamed Ahafhaf** (CLIPS-IMAG). Ce doctorat en sciences du langage concerne l'évaluation des systèmes de dialogue oral homme - machine. Il étudie l'utilisation, au niveau de la compréhension et de la gestion du dialogue, du paradigme d'évaluation DCR que j'avais proposé avec Jean Caelen (directeur de la thèse) et Jérôme Zeiliger. C'est à ce titre que Mohamed Ahafhaf a recherché mon concours. Après avoir travaillé avec moi sur l'évaluation des systèmes de compréhension, il s'intéresse désormais avec Jean Caelen aux niveaux de gestion du dialogue. Ce doctorat devrait donner lieu à une soutenance en 2004.

J'assure également depuis décembre 2002 le co-encadrement officiel de la thèse de **Véronique Bralé** qui se situe à l'interface entre l'informatique et les sciences du langage. Ce doctorat étudie l'utilisation de connaissances linguistiques permettant l'intégration de modes d'expressivité en synthèse de parole. Le cadre applicatif retenu concerne les systèmes de dialogue oral homme - machine finalisé. Ce doctorat de l'Université de Bretagne Sud est financé par France Télécom R&D. Il est essentiellement réalisé au sein de l'équipe synthèse de France Télécom R&D (Thierry Moudenc et Valérie Maffiolo) à Lannion. J'en assure l'encadrement académique en compagnie de Ioannis Kanellos (ENST Bretagne). Mon intervention consiste pour l'instant à l'évaluation des choix techniques ou méthodologiques effectués par Véronique Bralé.

Je rappellerai enfin d'autres encadrements de DEA en cours ou passés :

[9] Voir à ce sujet le modèle structurel de Chelba et Jelinek : Chelba C., Jelinek F. (2000) Structured language modeling. *Computer Speech and Language*, 14(4), 283-332.

[10] Schadle I. (2003) Sibylle : système linguistique d'aide à la communication pour les personnes handicapées. Doctorat Université de Bretagne Sud, Vannes, France. 18 décembre 2003. Rapport de recherche VALORIA-CORAIL-2003-03.

- *DEA de Frédéric Lamie* — Le stage de DEA de Frédéric Lamie est actuellement en cours de réalisation au sein du service informatique de l'aéroport de Brest Guipavas. Ce stage a une forte connotation R&D (Recherche et Développement). Son objectif est en effet de mettre en place un serveur vocal de démonstration pour le renseignement aérien en utilisant des technologies développées par TELISMA, une société issue d'un essaimage de France Telecom R&D (ex-CNET). J'assure l'encadrement scientifique de ce stage. La plate-forme commercialisée par TELISMA interdit toute initiative dans la conception des serveurs. Les aspects purement recherche de ce travail concerneront donc l'évaluation subjective du serveur vocal par des utilisateurs naïfs.
- *DEA de Laurent Derouard et Igor Schadle* — J'ai assuré au cours des années 1997 et 1998 l'encadrement de deux stages de DEA dans le cadre d'une collaboration avec Daniel Memmi, du laboratoire LEIBNIZ-IMAG. Cette collaboration portait sur l'utilisation de techniques connexionnistes (réseaux de neurones récurrents) pour le traitement du langage parlé. Plus spécifiquement, nous nous sommes intéressés au problème de la modélisation des variations de l'ordre canonique en français parlé. Si je ne développe pas de recherches dans le domaine des réseaux de neurones, nous verrons ultérieurement que cette question linguistique retient mon attention depuis plusieurs années. Si le stage de DEA de Laurent Derouard n'a pas donné les résultats espérés, Igor Schadle a repris avec succès cette problématique l'année suivante. Son travail a ainsi donné lieu à une communication dans une conférence internationale (*Eurospeech'99*).

3.2. Renforcer le potentiel de recherche local autour de l'interaction langagière

S'il ne rentre pas directement dans les préoccupations du groupe CORAIL, le suivi des doctorats de Mohamed Ahafhaf et de Véronique Bralé n'est cependant pas anecdotique. Il traduit en effet un effort d'ouverture pluridisciplinaire (sciences du langage) vers d'autres thématiques connexes à la communication homme - machine.

Cette ouverture est essentielle à mes yeux. Centré sur l'utilisateur, notre programme de recherche repose en effet sur une analyse des usages langagiers qui relève autant de la linguistique de corpus que de l'ingénierie des langues. Le laboratoire VALORIA s'étant construit autour d'une logique disciplinaire[11], il est malheureusement très difficile au groupe CORAIL de réunir en son sein la diversité de compétences qu'il peut exister dans d'autres centres de recherche.

Au cours de ces années passées à l'Université de Bretagne Sud, j'ai cherché à renforcer le potentiel des recherches autour du langage et de l'interaction homme - machine. Un de mes objectifs était précisément de renforcer la pluridisciplinarité de nos travaux. Par delà les relations informelles que j'ai pu nouer avec d'autres laboratoires locaux, je résumerai ici deux actions principales :

- ouverture d'un DEA en interaction homme - machine,
- proposition de création d'un pôle en linguistique de corpus à l'Université de Bretagne Sud (plan pluri-formation (PPF) entre les laboratoires ADICORE et VALORIA en négociation dans le cadre du prochain contrat quadriennal de l'université).

On peut juger que la dimension purement scientifique de ces activités est limitée. Elles n'en sont pas moins essentielles à la pérennité des recherches du groupe CORAIL.

Lancement du DEA Informatique mention « Interaction Homme - machine » — Jusqu'en 2002, le laboratoire VALORIA n'était rattaché à aucun DEA. S'il était reconnu

[11] Le VALORIA se définit en effet comme le laboratoire d'informatique de l'Université de Bretagne Sud (VAnnes LOrient Recherches en Informatique et ses Applications).

comme laboratoire d'accueil du DEA informatique de l'Université de Rennes 1, il était difficile d'attirer de jeunes chercheurs sur le site de Vannes. Comme bien souvent dans le cas de DEA disciplinaires, ces étudiants en informatique n'ont bénéficié en outre que d'une ouverture limitée vers l'interaction homme - machine ou le traitement des langues naturelles.

Franck Poirier et moi-même entretenant des relations scientifiques suivies avec le département IASC (Intelligence Artificielle et Sciences Cognitives) de l'ENST Bretagne, nous avons saisi l'opportunité de lancer en commun un DEA en informatique localisé géographiquement sur la Bretagne et centré sur la thématique pluridisciplinaire de l'interaction homme - machine. Nous avons travaillé pendant plus de deux ans sur l'ouverture de ce DEA co-habilité par l'Université de Bretagne Sud (sceau principal) et l'ENST Bretagne. Cette formation dirigée par Franck Poirier a accueilli sa première promotion au cours de l'année universitaire 2002-2003.

Après trois mois de tronc commun, chaque site accueille des enseignements spécifiques donnant une coloration thématique au cursus. Au terme de sa formation, un étudiant ayant choisi le site de Vannes aura ainsi bénéficié d'enseignements sur[12] :

- l'interaction et la coopération homme - machine,
- l'aide à la décision,
- la communication langagière et l'ingénierie des langues,
- l'interaction gestuelle et l'animation d'images virtuelles,
- la robotique,
- les méthodes d'apprentissage utilisées dans ces différents thématiques.

J'assure l'intégralité des enseignements en linguistique, traitement automatique des langues naturelles et communication homme - machine. Je suis par ailleurs co-ordonnateur pédagogique du DEA sur le site de Vannes.

Une douzaine d'étudiants ont suivi durant cette année les enseignements du DEA sur le site de Vannes. En dépit de difficultés récurrentes de financement de thèses, ce flux d'étudiants de troisième cycle devrait permettre une consolidation des recherches réalisées au VALORIA.

Mise en place d'un pôle « linguistique corpus » à l'Université de Bretagne Sud — Comme je l'ai relevé précédemment, l'absence de linguistes ou de philosophes du langage au sein du laboratoire VALORIA est un frein à la conduite des recherches du groupe CORAIL. C'est pourquoi je me suis rapproché de Geoffrey Williams et ses collègues, qui mènent des travaux en linguistique de corpus au sein du laboratoire ADICORE (ex-CRELLIC, Lorient) de l'Université de Bretagne Sud.

Les travaux d'ADICORE, qui ont une finalité terminologique, restent éloignés de nos centres d'intérêts. Cependant, les pratiques scientifiques de nos deux équipes reposent semblablement sur l'analyse d'observations langagières extraites de corpus réels. C'est pourquoi j'ai proposé à Geoffrey Williams de participer à la mise en place d'un pôle scientifique local autour de la linguistique de corpus.

Cette collaboration s'est tout d'abord traduite par la co-organisation d'un colloque annuel en linguistique de corpus (LINGCORP) qui est désormais bien reconnu par la communauté scientifique. Sous mon pilotage, elle repose désormais sur la proposition d'un plan pluri-formations (PPF CORAIL) qui vise la mise en place d'un centre de ressources en linguistique de corpus. Cette structure sera dédiée à la constitution et à la diffusion de corpus d'envergure pour l'ingénierie des langues et la linguistique appliquée. Les ressources linguistiques réalisées, qui devraient représenter un corpus diversifié de 500 000 mots pour la partie orale,

[12] Par un jeu d'option limité, le cursus choisit par un étudiant peut-être légèrement moins étendu. L'emploi du temps du DEA permet néanmoins à chaque étudiant de suivre l'ensemble des options s'il le souhaite.

seront mises librement à la disposition de la communauté scientifique. Au terme du PPF, notre objectif est de pérenniser ce pôle sous la forme d'une Equipe de Recherche Technologique (ERT) consacrée exclusivement à la création et la diffusion de ressources linguistiques libres.

L'autre ambition du projet est bien entendu de favoriser les convergences entre des communautés relevant des sciences du langage et de l'informatique appliquée à la communication homme - machine. On peut en effet espérer que ce PPF favorisera l'échange de connaissances, de savoir-faire ainsi que l'émergence d'intérêts scientifiques communs entre les chercheurs des deux laboratoires.

Cette demande est actuellement en cours d'évaluation dans le cadre de la validation du contrat quadriennal de l'Université de Bretagne Sud (vague 2004 de contractualisation). S'il est accepté, ce projet donnera une forte visibilité aux recherches du groupe CORAIL. C'est là le second objectif de mes activités au sein de l'Université de Bretagne Sud.

4. ANCRER NOS ACTIVITES AU SEIN DE LA COMMUNAUTE

C'est un lieu commun que d'affirmer que la recherche technologique se réalise en réseau. Par delà l'intérêt des échanges scientifiques, les collaborations sont désormais essentielles à la mise en œuvre de projets ou de campagne d'évaluation d'envergure sans lesquels il ne serait pas possible de parler d'ingénierie des langues. Lorsque, jeune maître de conférences, vous êtes amené à créer un nouveau groupe de recherche en dehors d'un pôle d'excellence scientifique, la question de l'isolement de cette équipe se pose avec une acuité accrue. C'est pourquoi j'ai constamment cherché à ancrer et à faire reconnaître nos activités au sein de la communauté scientifique.

Ce paragraphe fait le bilan de ces actions, en partant des activités relevant de l'animation de la recherche (organisation de congrès, participation à des structures d'animation scientifique) pour finir avec l'ensemble des collaborations scientifiques réalisées au cours de ces sept dernières années.

4.1. Organisation de manifestations scientifiques

J'ai participé, généralement avec l'aide de mes collègues du groupe CORAIL et d'EQUIPAGE, à l'organisation de plusieurs manifestations scientifiques dans des domaines qui couvrent l'ensemble des centres d'intérêts de notre équipe. Ces activités relevant avant tout de l'animation de la recherche, je vais les évoquer très brièvement.

Rencontres Jeunes Chercheurs en Interaction Homme - machine (RJC-IHM 2000) — Lancées à l'initiative de Franck Poirier et moi-même, ces rencontres jeunes chercheurs se sont déroulées sur l'île de Berder, dans le golfe du Morbihan, sous le patronage scientifique de l'Association Française d'Interaction Homme - machine (AFIHM). J'ai assuré la présidence du comité d'organisation de ce colloque qui a réunit une trentaine de doctorants et chercheurs confirmés du domaine.

Journées de la Linguistique de Corpus (LINGCORP 2001, 2002 et 2003) — Ce congrès connaîtra cette année sa troisième édition. Il est organisé chaque année à Lorient à l'initiative de Geoffrey Williams (ADICORE). Sa création répondait à l'absence de congrès francophone dans ce domaine, ainsi qu'au désir de Geoffrey et moi-même de rapprocher les recherches en linguistique appliquée et en ingénierie des langues. Si ce dernier objectif n'a été qu'en partie atteint à mon sens, LINGCORP connaît une audience croissante d'année en année. J'en assure la co-organisation depuis la première édition.

Cette collaboration entre les laboratoires ADICORE et VALORIA se poursuivra par ma participation au comité d'organisation du congrès EURALEX'2004, qui se déroulera à Lorient du 6 au 10 juillet 2004.

Autres manifestations — J'ai par ailleurs participé aux comités d'organisation d'autres congrès parmi lesquels TALN'2003, organisé à Batz-sur-Mer par l'IRIN en collaboration avec l'IRISA et le VALORIA. Je ne détaillerai pas ici l'organisation de deux ateliers thématiques dans le cadre des conférences TALN. Ceux-ci répondaient en effet avant tout à des motivations scientifiques que je détaillerai plus loin (cf. § 5.1 et 5.2).

4.2. Participation à des structures d'animation de la recherche

La reconnaissance de travaux du groupe CORAIL s'est également traduite par ma participation à diverses structures d'animation de la recherche. Je citerai en particulier ma contribution aux travaux des comités suivants :

- **comité directeur du GDR I3 (Intelligence - Information - Interaction) du CNRS** — J'ai été appelé à rejoindre en 2002 le comité directeur du GDR suite à la direction que j'assure du groupe de travail (GT 5.5) sur la compréhension de parole. Je présentai plus loin (cf. § 5.1) les aspects purement scientifiques de cette activité d'animation.
- **conseil d'administration de l'ATALA** — J'ai rejoint en 2002 le conseil d'administration de l'Association pour le Traitement Automatique des LAngues. Cette société savante joue un rôle central dans l'animation scientifique de la communauté francophone en ingénierie des langues.

Il est enfin une autre forme d'animation scientifique sur laquelle je souhaite insister, à la fois parce qu'elle est très prenante mais également parce qu'elle me tient particulièrement à cœur. Il s'agit de la direction de la revue *In Cognito — Cahiers Romans de Sciences Cognitives* que j'ai créée en 1995 et dont je suis toujours le rédacteur en chef.

Direction des *Cahiers Romans de Sciences Cognitives* — Comme on a pu le noter, ma recherche concerne des champs disciplinaire variés regroupant aussi bien l'Intelligence Artificielle, le traitement automatique du langage (TALN et TALP), l'interaction homme - machine que les sciences du langage (linguistique de corpus essentiellement). Cette approche interdisciplinaire, associée à la primauté que j'accorde à la prise en compte des usages et besoins réels de l'usager, entre en résonance avec le programme des Sciences Cognitives pour lequel j'ai toujours marqué un grand intérêt.

C'est ainsi que j'avais suivi en auditeur libre les enseignements du DEA Sciences Cognitives grenoblois lors de ma première année de doctorat. Je participe depuis cette époque aux activités de l'association *In Cognito* des chercheurs rhônalpins en sciences cognitives, dont je suis actuellement le président. C'est dans le cadre de cette association que j'ai proposé en 1996 la création d'une revue scientifique éponyme destinée à soutenir les recherches francophones du domaine.

Le lancement de cette revue répondait alors à un besoin fort, du fait de l'absence d'espace de diffusion francophone consacré aux recherches novatrices en sciences cognitives. La seule revue francophone du domaine, *Intellectica*, publiait en effet essentiellement des articles suscités auprès de chercheurs reconnus. En négligeant les soumissions spontanées, cette publication de qualité délaissait les recherches émergentes les plus originales.

Dès sa création, *In Cognito* chercha à combler ce vide sous ma direction. La tâche ne fut pas aisée, tout d'abord parce que la revue ne s'adosse pas à un éditeur privé, mais également parce qu'elle fut perçue initialement comme une publication destinée aux jeunes chercheurs.

In Cognito a cependant su faire sa place au cours du temps. A ma connaissance, notre diffusion égale désormais celle de la revue *Intellectica*. Par ailleurs, nos auteurs sont désormais aussi bien des scientifiques reconnus que des jeunes chercheurs dont nous soutenons toujours les travaux les plus originaux.

Forts de ce succès, nous avons décidé de donner une audience encore plus importante à la revue. Grâce au soutien financier du Pôle Rhône-Alpes en Sciences Cognitives, dirigé par Nicolas Balacheff et Hélène Paugham - Moisy, *In Cognito* est en effet depuis 2003 une publication internationale éditée en quatre langues romanes sous le titre (en français) des *Cahiers Romans de Sciences Cognitives*[13]. Cette montée en puissance parachève notre objectif initial de reconnaissance de la publication scientifique en français. Les *Cahiers Romans de Sciences Cognitives* offrent en effet aux chercheurs francophones — mais aussi « romans » — l'opportunité de valoriser leur travaux par une publication en langue maternelle qui s'adressera à des scientifiques d'Europe, d'Amérique latine ou du Nord (figure 0.1).

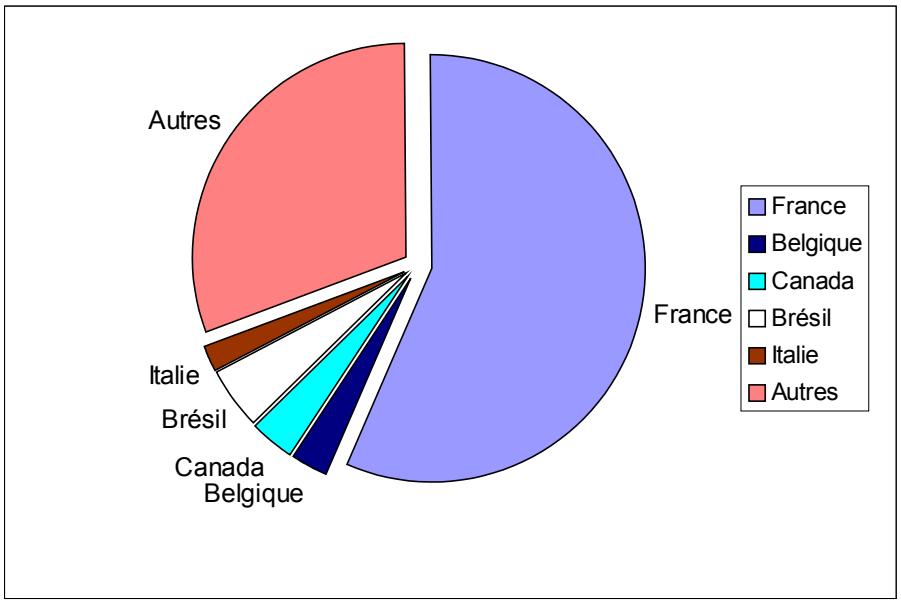

Figure 0.1. *— Répartition géographique des abonnés aux Cahiers Romans de Sciences Cognitives (01/12/2003).*

Cette revue internationale est désormais dirigée par quatre rédacteurs en chef : Plinio Barbosa (UNICAMP, Campinas, Brésil), Tersa Inès Ceratto (U. Stockholm), Maria Poala d'Imperio (LPL, Aix-en-Provence, France) et moi-même. Elle est référencée par les *Cambridge Scientific Abstracts* ainsi que dans la base *E-psyche*.

Si cette aventure fut et reste très prenante en terme de charge de travail, elle m'a également permis de nouer certaines collaborations. Ce fut en particulier le cas de mes premiers échanges, autour d'un numéro spécial publié sous la direction de Philippe Lenca et ses collègues du département IASC de l'ENST Bretagne.

Je vais précisément détailler maintenant les collaborations les plus importantes auxquelles j'ai participé depuis ma nomination.

[13] Toile : http://www.in-cognito.net

5. DES COLLABORATIONS POUR UN PROJET SCIENTIFIQUE

La majeure partie des collaborations que je vais présenter ci-dessous ont été initiées sous ma direction. Ces collaborations particulières ont un objectif qui dépasse la « simple » intégration du groupe CORAIL au sein de la communauté scientifique. Elles visent en effet à favoriser une réflexion globale sur des thématiques (ressources linguistiques, évaluation diagnostique des systèmes) ou des approches (méthodes hybrides TALN/TALP) qui sont au cœur de mon projet scientifique et qui demandaient selon moi à être plus présentes dans les débats qui animent l'ingénierie des langues. Je vais présenter ces collaborations suivant les différents axes de recherche qui structurent les travaux du groupe CORAIL.

5.1. Dialogue Homme - machine

Action de Recherche Concertée « Dialogue Oral » de l'AUF — La première collaboration à laquelle a participé le groupe CORAIL faisait suite à mes travaux grenoblois (CLIPS-IMAG) au sein de l'action de recherche concertée « Dialogue oral » (ARC ILOR B2) de l'Agence Universitaire de la Francophonie (AUF, ex-AUPELF-UREF). Du fait de problèmes de financement récurrents, cette action n'a pas pu tenir tous ses objectifs. Elle s'est limitée à une réflexion sur les méthodologies d'évaluation des systèmes de dialogue ainsi qu'à la constitution de corpus oraux. Cette seconde activité sera présentée dans le paragraphe 5.3.

A la suite des programmes d'évaluation financés par la (D)ARPA américaine, le propos initial de cette action de recherche concernait l'évaluation des systèmes de compréhension de parole et de dialogue oral. Très vite, deux approches de test ont émergé au sein des participants (CLIPS-IMAG, ICP, IRISA/CORDIAL, IRIT, LIMSI, VALORIA). D'un côté, certains laboratoires tenaient à une évaluation quantitative globale suivant un paradigme proche de ceux utilisés dans les campagnes de la (D)ARPA sur le renseignement aérien (projet ATIS). Sans nier l'intérêt de ce type d'évaluation objective, les autres participants regrettaient son manque de pouvoir diagnostic (voir le chapitre 3, consacré à l'évaluation pour une discussion plus approfondie). C'est ainsi que j'ai proposé avec Jean Caelen (CLIPS-IMAG) et Jérôme Zeiliger (ICP) une nouvelle méthodologie d'évaluation appelée DQR (Demande - Question - Réponse). Inspirée de certains travaux issus du TALN (FraCas, TSNLP), ce paradigme repose sur la définition de multiples jeux de tests spécifiques à l'étude de classes particulières de difficultés (procédés de l'oral spontané, références anaphoriques, gestion des erreurs de reconnaissance, etc.). Présentée aux participants de l'ARC, la méthodologie a ensuite connu certaines améliorations auxquelles a également contribué Jacques Siroux (IRISA/CORDIAL). Reconnue désormais sous le nom d'évaluation DCR (Demande – Contrôle - Référence), cette proposition fait l'objet d'une réflexion sur son application aux niveaux dialogiques dans le cadre du doctorat de Mohamed Ahafhaf. Elle a donné lieu à différentes publications nationales et internationales (revue *Langues*, journées *JST-Francil*, conférences *LREC*).

L'arrêt du soutien financier de l'AUF à ces actions de recherche a empêché la mise en œuvre d'une campagne de tests reposant sur ce paradigme. La méthodologie DCR a cependant eu une influence sensible sur les recherches francophones en matière d'évaluation. Elle a ainsi été reprise sous une forme allégée dans le cadre de la campagne d'évaluation par DEFI du GDR I3. De même, elle a clairement inspiré la méthodologie d'évaluation PEACE qui sera utilisée dans le cadre de la campagne de test MEDIA du programme TECHNOLANGUE. Le groupe CORAIL collabore à ces deux projets que je vais maintenant présenter.

Groupe de travail « Compréhension robuste de la parole » du GDR I3 — J'ai créé ce groupe de travail dès le lancement du GDR I3 (Intelligence – Information – Interaction) par le CNRS en 1998. A l'origine, nos travaux s'intégraient à ceux d'un groupe de travail plus général (GT « Parole » dirigé par Régine André-Obrecht) et ne se limitaient pas à l'évaluation

des systèmes. Suivant une préoccupation qui guide mon projet de recherche, mon propos était en effet de favoriser le rapprochement des techniques utilisées respectivement en traitement du langage (TALN, langage écrit) et de la parole (TALP). Alors qu'elles travaillent sur des objets très proches, ces deux communautés ont longtemps vécu dans une relative ignorance mutuelle. Poursuivant des objectifs sensiblement différents, elles ont généralement privilégié l'utilisation de techniques différentes : formalismes linguistiques d'une part et modèles stochastiques de langage d'autre part.

Au tournant des années 1990, l'émergence d'une démarche ingénierique en TALN, de même que l'observation expérimentale des limitations des modèles probabilistes par le TALP, ont cependant posé la question de l'utilisation conjointe des techniques issues des deux communautés. C'est ainsi que le TALP a commencé à s'interroger sur l'intégration de modèles linguistiques plus fouillés[14].

Alors que le CNRS lançait le nouveau GDR I3, il n'existait pourtant en France aucun programme de recherche, aucune structure d'animation s'intéressant à cette question. Un des rôles des GDR est d'animer la communauté scientifique autour de sujets émergents et fédérateurs. Aussi ai-je proposé la mise en place d'un groupe de travail destiné à favoriser les échanges entre le TALN et le TALP. Plusieurs réunions de travail ont été organisées dans ce cadre au cours des années 1998 et 1999. L'une de ces journées a pris la forme d'un **atelier thématique « méthodes hybrides TALN/TALP »** que j'ai organisé avec Damien Genthial (CLIPS-IMAG) dans le cadre du congrès TALN'1999[15]. Ces différentes journées étaient centrées sur l'exposé des travaux des différents participants. Elles ont favorisé une reconnaissance mutuelle des travaux de chaque communauté.

Au bout de deux ans d'existence, ce groupe de travail n'avait cependant répondu que partiellement à mes attentes. Regroupant des chercheurs travaillant sur des thématiques très diverses et doté d'un budget limité, il pouvait difficilement permettre un effort de recherche plus profond dépassant la simple animation scientifique.

J'ai donc décidé de recentrer les activités du groupe sur la thématique qui y était la plus représentée et qui constitue un des cadres applicatifs étudiés par CORAIL : la compréhension de la parole. Dès lors, le groupe a réuni l'ensemble des laboratoires français (CLIPS-IMAG, IRIT, LIMSI, LORIA, VALORIA) travaillant sur cette problématique. Il constitue depuis 2002 un groupe de travail à part entière, sous l'intitulé « Compréhension robuste de la parole » (GT 5.5). Je le représente désormais au sein du comité directeur du GDR I3. Depuis cette réorientation, le groupe a connu une activité soutenue, pour deux raisons principales :

- les laboratoires impliqués sont tous intéressés par la problématique de l'intégration de techniques issues du TALN. Certains s'inscrivent en effet déjà dans une telle perspective tandis que d'autres ont pris conscience des limites des méthodes probabilistes, issues du TALP, qu'ils utilisent actuellement.
- les activités du groupe ont coïncidé avec le déroulement de plusieurs doctorats sur la compréhension de parole[16]. Le groupe de travail a ainsi pu bénéficier de l'investissement

[14] On citera, à titre d'exemple parmi d'autres, les travaux du MIT autour du système TIAN : Seneff S. (1992) TINA: a natural language system for spoken language applications. *Computational Linguistics*, 18(1). 61-86

[15] Antoine J.-Y., Genthial D. (1999), *Méthodes hybrides issues du TALN et du TAL Parlé : états des lieux et perspectives*, TALN'1999, atelier thématique « Méthodes hybrides pour le TAL robuste », Cargèse, France, 1-17.

[16] Doctorats de Sophie Rosset (LIMSI, U. Paris XI, décembre 2000), Caroline Bousquet (IRIT, U. Paul Sabatier, septembre 2002), Jérôme Goulian (VALORIA, U. Bretagne Sud, décembre 2002), Mohamed -

de participants désireux de profiter de ces échanges.

C'est dans ce cadre que les activités du groupe ont été orientées vers l'évaluation des systèmes. Il est en effet rapidement apparu qu'une réflexion approfondie sur nos différentes approches ne pouvait être menée à bien sans disposer de renseignements sur le comportement des systèmes concernés. Pour les participants, cette évaluation devait avoir une portée diagnostique précise. Le budget du groupe de travail étant essentiellement consacré au financement de journées d'études, cette campagne de test se devait cependant d'être légère.

Pour répondre à ces besoins contradictoires, j'ai proposé un nouveau paradigme de test (évaluation par défi) qui est adapté de la méthodologie DCR et repose sur des séries de tests spécifiques à un ensemble de phénomènes. Ce paradigme a été utilisé dans le cadre d'une première campagne d'évaluation par défi. Les résultats de la campagne ont été débattus au cours d'une journée d'étude à Toulouse. Ils ont donné lieu à une publication commune au congrès LREC'2002 ainsi qu'à d'autres publications personnelles des participants.

Cette première campagne de test a fourni un diagnostic intéressant des apports et limitations des techniques utilisées par chacun. Pour pouvoir servir de cadre à nos réflexions futures sur l'évaluation de la compréhension de parole, il importe cependant que ce type d'évaluation s'intègre dans un cadre méthodologique plus précis. C'est pourquoi le groupe travaille actuellement sur un recensement exhaustif des problèmes — traitement de phénomènes linguistiques particuliers, gestion des problèmes techniques tels que les erreurs de reconnaissance, etc.— qui se posent actuellement à la compréhension de parole. La feuille de route qui résultera de cette étude sera certainement utile aux activités du projet MEDIA, qui constitue une suite naturelle des activités du groupe.

Evaluation en contexte de la compréhension de la parole : **projet MEDIA**— Proposé par le laboratoire LIMSI (Laurence Devillers), ce projet s'intègre au programme d'évaluation EVALDA organisé dans le cadre de l'action TECHNOLANGUE du Ministère de la Recherche. Il a été construit autour des activités des participants au GT 5.5 du GDR I3 et dispose de financements permettant une évaluation à grande échelle des systèmes de compréhension de parole. Cette campagne de test ne se limitera pas à la compréhension « littérale » des énoncés mais s'intéressera également à leur interprétation en contexte (calcul des références, par exemple), ce qui constitue une grande originalité dans le domaine. L'évaluation reposera sur le paradigme de test PEACE proposé par le LIMSI. En s'intéressant à l'étude du comportement détaillé des systèmes, cette méthodologie constitue une tentative de rapprochement des approches globales de type (D)ARPA ATIS et des évaluations diagnostique du genre de DCR ou de DEFI, dont elle s'inspire.

Les deux systèmes de compréhension développés au VALORIA seront évalués dans le cadre de cette campagne. Un des objectifs du DEA de Julien Foulon, co-encadré par Jérôme Goulian et moi-même, est précisément d'adapter le système ROMUS à la partie « interprétation » de MEDIA. Pour l'heure, les participants au projet travaillent sur la constitution des corpus oraux sur lesquels se dérouleront les tests. Ces corpus seront recueillis par la technique du magicien d'Oz, le cadre applicatif retenu étant le renseignement touristique. Cette campagne de test prendra fin en 2004.

Comme je le montrerai par la suite, l'utilisation de méthodes fouillées issues du TALN, de même que la recherche d'une évaluation à fort pouvoir diagnostic, sont au cœur de mon projet

Zakaria Kurdi (CLIPS-IMAG, U. Joseph Fourier, avril 2003), Jeanne Villaneau (VALORIA, U. Bretagne Sud, en cours). Salma Jamoussi (LORIA, doctorat de l'U. Henri Poincaré) et Mohamed Ahafhaf (CLIPS-IMAG, doctorat de l'U. Stendhal) ont rejoint plus récemment le groupe. Je tiens à remercier ces jeunes docteurs pour leur implication dans les travaux du groupe, de même que les autres participants : Nadine Vigouroux (IRIT) et Laurence Devillers (LIMSI).

scientifique. Ces collaborations m'ont donc permis d'amener ces problématiques émergentes parmi les réflexions des chercheurs en dialogue oral homme - machine.

A un degré moindre, mes collaborations sur le handicap visent elles aussi à porter au sein de la communauté scientifique les préoccupations du groupe CORAIL.

5.2. Dialogue médié par l'ordinateur : aide au handicap

Alors qu'il répond à un besoin social fort, l'aide au handicap est une thématique de recherche relativement délaissée qui n'a bénéficié, jusqu'à une date récente que de peu de structuration institutionnelle[17]. Si la robotique la reconnaît comme un cadre applicatif significatif, elle ne concerne au contraire qu'à la marge l'ingénierie des langues[18]. L'aide à la communication langagière est pourtant aussi importante que l'aide aux gestes de la vie quotidienne dans la quête d'autonomie des personnes handicapées.

Cette situation est un frein indéniable aux recherches du domaine. C'est pourquoi il m'apparaît nécessaire de structurer la communauté afin de permettre au minimum une capitalisation des expériences de chacun. A titre d'exemple, il est ainsi regrettable que les différents systèmes francophones d'aides à la saisie qui ont été réalisés au cours de la dernière décennie n'aient jamais été l'objet d'une campagne d'évaluation commune.

Atelier thématique « Handicap » — Pour répondre à cette insuffisance, j'ai tout d'abord organisé avec Brigitte Le Pévédic un atelier thématique consacré à l'ingénierie des langues pour le handicap, dans le cadre du congrès TALN'2001[19]. Cet atelier a répondu de manière évidente à un besoin fort, puisque la quasi-totalité des laboratoires francophones concernés par le sujet y furent représentés. Les échanges se sont concentrés sur une thématique qui réunit la plupart des efforts de la communauté, à savoir l'aide à la saisie de texte pour la communication (orale ou écrite) des personnes fortement handicapées. Les systèmes qui ont été présentés reposent sur une très grande diversité d'approche. Celle-ci concerne aussi bien les techniques d'analyse utilisées (formalismes linguistiques, grammaires probabilistes ou simples modèles de langage markoviens) que le cadre dans lequel est envisagé l'aide au handicapés (désabréviation, prédiction linguistique libre, co-génération de texte, normalisation à visée ré-éducatrice pour patients aphasiques).

Cette diversité est une richesse pour les recherches futures du domaine. Il est cependant indispensable qu'elle s'accompagne d'une validation sérieuse des différentes approches envisagées pour permettre de réelles avancées. En parallèle aux activités « Handicap » de l'ACI Cognitique ou de l'IFRATH[20] , le lancement du RTP Handicap[21] par le CNRS (département STIC) offre un cadre institutionnel idéal pour cette confrontation des idées et des résultats. C'est pourquoi je travaille actuellement avec Nadine Vigouroux, co-animatrice du RTP Handicap, à la mise en place d'un Equipe Projet consacrée à l'évaluation des systèmes francophones d'aide à la saisie de texte pour handicapés.

Généralisation des recherches en handicap à l'ingénierie des langues : programme EASy d'évaluation des analyseurs syntaxiques — En organisant l'atelier « Handicap » au congrès TALN'2001, je m'étais assigné un second objectif que je rappelais dans le texte

[17] Ce domaine de recherche dispose pourtant de soutiens financiers appréciables, en particulier dans le cadre des programmes européens TIDE (*Technology Initiatives for Disabled and Elderly people*).

[18] En France, l'aide au handicap ne concerne ainsi que des équipes très restreintes dans quelques laboratoire en ingénierie des langues (ENST Bretagne, IRIT, IRIN et désormais LI Tours, LIM, LPL et VALORIA).

[19] Antoine J.-Y, Le Pévédic B. (2001), Ingénierie des Langues et Handicap, actes *TALN'2001*, atelier thématique « Handicap et Ingénierie Linguistique », Tours, France

[20] Toile : http://www.lasc.univ-metz.fr/ifrath

[21] Toile : http://www.irit.fr/RTP-Handicap/

d'introduction aux communications présentées[22]. Il s'agissait de faire (re)connaître l'aide au handicap comme une thématique d'importance de l'ingénierie des langues, que ce soit d'un point de vue théorique ou comme cadre applicatif. De mon point de vue, cette tentative ne fut cependant réussie qu'à moitié.

Cette reconnaissance mutuelle me semble pourtant essentielle. D'une part parce que l'aide linguistique au handicap, faiblement représentée au sein de la communauté scientifique, doit s'appuyer plus fortement sur les avancées de l'ingénierie des langues. Mais aussi, en contrepartie, parce que l'aide au handicap répond à des contraintes et des motivations fortes (robustesse d'analyse, personnalisation et adaptation des systèmes, prise en compte crucial de l'utilisateur) qui sont au centre des préoccupations actuelles de l'ingénierie des langues.

A mon sens, ce rapprochement ne sera effectif que lorsque l'aide au handicap aura fait montre de son apport à l'ingénierie des langues par ses pratiques méthodologiques et ses résultats. C'est pourquoi le lancement rapide d'un programme d'évaluation des systèmes d'aide à la saisie de texte me semble être une priorité.

En attendant cette action commune, le groupe CORAIL participe déjà à cet effort de reconnaissance dans le cadre du programme EASy d'évaluation des analyseurs syntaxiques du français (projet EVALDA de l'action TECHNOLANGUE). Comme je l'ai évoqué précédemment, le système Sibylle réalisé dans le cadre du doctorat d'Igor Schadle repose sur un modèle qui est proche des tentatives récentes d'introduire des connaissances linguistiques plus profondes dans les modèles de langage probabilistes. Son intérêt dépasse donc le « simple » cadre de l'aide au handicap pour concerner d'une manière générale l'ingénierie des langues. Aussi ai-je proposé que notre système participe à ce programme qui constitue la première campagne de test d'envergure jamais réalisée sur des analyseurs syntaxiques francophones. Cette participation permettra de comparer SIBYLLE avec des systèmes à portée plus générale reposant sur des techniques alternatives à la notre (grammaires formelles éventuellement probabilisées). Le projet EASy représente donc une opportunité intéressante pour généraliser la portée de nos travaux sur le handicap dans un cadre de réflexion se situant à l'interface des modèles linguistiques fouillés et des approches stochastiques.

EASy reposera sur deux campagnes successives de test. Tout d'abord, on évaluera les capacités des systèmes à segmenter les énoncés — oraux ou écrits — de tests en constituants minimaux non récursifs (*chunks*). Ensuite, l'évaluation portera sur l'élaboration des structures syntaxiques proprement dites. Celles-ci seront envisagées comme des relations de dépendances syntaxiques (sujet, objet, etc.) entre les *chunks* issus de la segmentation précédente. Le système SIBYLLE n'a pas été conçu pour typer les dépendances entre les chunks qu'il a caractérisés. Comme d'autres systèmes participants, il ne sera donc évalué que sur la première étape de segmentation. Cette campagne de test prendra fin en 2004.

Handicap et utilisateurs : collaboration avec Kerpape — Les collaborations que j'ai présentées ci-dessus restent centrées autour de l'ingénierie des langues. L'aide au handicap requiert pourtant des études ergonomiques pour s'assurer que les systèmes d'assistance répondent aux besoins et capacités de chaque patient. On sait en effet que le meilleur système de prédiction linguistique ne sera d'aucune utilité s'il présente des difficultés rédhibitoires d'utilisation. C'est dans cet esprit que j'ai invité Jean-Marc Toulotte et Brigitte Cantegrit (I3D, U. Lille 1 et Institut Régional de Recherche sur le Handicap) à participer à l'atelier « handicap » de TALN'2001. De même, nos travaux sur le système SIBYLLE s'effectuent en collaboration étroite avec le Centre de Réadaptation Fonctionnelle de Kerpape. Ces échanges concernent aussi bien l'analyse amont des besoins des utilisateurs que la validation aval de nos prototypes auprès des personnes handicapées du centre. Portant en particulier sur la

[22] Antoine J.-Y, Le Pévédic B. (2001), *op. cit*.

conception de l'interface homme - machine du système SIBYLLE, cette collaboration s'effectue en premier lieu avec Jean-Paul Departe et Alain Scaviner, du laboratoire d'informatique et d'électronique de Kerpape. Suivant les besoins, elle donne également lieu à des échanges très instructifs avec les thérapeutes ou les éducateurs du centre.

Cette collaboration rappelle le caractère pluridisciplinaire de l'approche centrée sur l'utilisateur qui guide mes travaux. C'est cette dimension qui m'a incité à orienter une partie des activités du groupe CORAIL vers la linguistique de corpus et la constitution de ressources linguistiques orales. Cet axe de recherche se retrouve dans les collaborations présentées ci-dessous.

5.3. Ressources linguistiques pour le dialogue oral

Le développement de nos activités en matière de constitution de ressources linguistiques est parti d'un constat simple. D'une part, la conception de systèmes interactifs centrée sur l'utilisateur nécessite une réflexion amont sur les usages langagiers des utilisateurs. Ces études doivent reposer de manière privilégiée sur des corpus réels, appelés *corpus pilotes* par Jean Caelen et ses collègues[23]. D'autre part, il n'existe pour l'heure que très peu de corpus de dialogue oral mis à la disposition de la communauté francophone. Afin de développer de manière autonome notre programme de recherche, je me suis engagé dans une politique de constitution de corpus destinée aux besoins du dialogue oral homme - machine. Les principales ressources qui ont été réalisées à l'heure actuelle ont bénéficié d'un soutien financier dans le cadre de deux collaborations[24].

ARC « Dialogue Oral » de l'AUF — Lors de la première phase de cette action de recherche (cf. § 5.1), le laboratoire CLIPS-IMAG a recueilli un corpus pilote de renseignement touristique enregistré à l'office du tourisme de Grenoble (corpus OTG). Distribué sous la forme d'enregistrements audio accompagnés de fichiers de description sommaire de chaque transaction recueillie, ce corpus intéressant était difficilement utilisable en l'état. J'ai donc proposé que le groupe CORAIL réalise sa transcription au cours de la seconde phase du projet. Le corpus OTG a été enregistré en situation réelle, donc dans des conditions audio relativement difficiles. C'est pourquoi il n'a été possible de transcrire que la partie la plus audible du corpus (25700 mots environ). Je compte poursuivre le développement de cette ressource à la fois dans le cadre du projet OURAL décrit ci-après (annotation morphosyntaxique) et sur fonds propres. C'est ainsi que les co-références anaphoriques et les réparations (répétitions, corrections) présentes dans le corpus seront annotées au cours des stage de DEA de Julien Foulon et Iadaloharivola Randria.

Vers une banque de corpus de dialogue oral : projet OURAL — Notre programme de constitution de corpus se poursuit désormais dans le cadre du projet OURAL (sous-projet du programme AGILE de l'action TECHNOLANGUE). Ce projet, qui réunit une quinzaine de participants issus d'horizons variés (TALN, TALP, recherche d'information, lexicographie mais aussi psycholinguistique), a pour finalité la mise en œuvre d'outils de base pour l'ingénierie des langues. Ces outils (étiqueteur en parties du discours, segmenteur, analyseur de mots inconnus, etc.) reposeront aussi bien sur une modélisation formelle que probabiliste.

---

[23] Caelen J., Zeiliger J., Bessac M., Siroux J., Perennou G. (1997) Les corpus pour l'évaluation du dialogue homme - machine. Actes des *1ères Journées Scientifiques et Techniques FRANCIL*, *JST'1997*, Avignon, France, 215-222. Texte repris dans : Chibout K., Mariani J., Masson N., Néel F. (Dir.) (2000) Ressources et évaluations en ingénierie des langues. De Boeck Université, Duculot, Bruxelle, Belgique. 417-435

[24] Sabine Letellier a par ailleurs constitué sur fond propre un petit corpus pilote (5300 mots) mettant en jeu des enfants autour d'une activité simulée de planification de loisirs et de renseignement touristique (corpus MASSY).

La réalisation et la validation de ces outils nécessitent la constitution d'importantes ressources linguistiques. Dans l'esprit de l'initiateur du projet, Claude de Loupy (SINEQUA), ces outils devaient uniquement concerner le traitement du langage écrit. Il est cependant manifeste que ces outils seront de plus en plus recherchés par le TALP au fur et à mesure qu'il s'orientera vers des traitements plus fins. Ainsi, l'étiquetage (semi)automatique des corpus oraux constitue désormais une question d'actualité[25]. C'est pourquoi j'ai proposé d'étendre la portée du projet en validant certains outils sur des corpus de parole spontanée. Deux types de ressources orales seront ainsi constituées au cours du projet :

- conversations radiodiffusées : corpus réalisé par le laboratoire SILEX[26].
- interaction orale homme - homme en situation de dialogue finalisé : corpus réalisé par le groupe CORAIL sous ma direction[27].

Ces corpus seront transcrits et recevront une annotation en parties du discours ainsi qu'une délimitation des entités nommées. Au terme du projet, CORAIL proposera une banque de corpus de dialogue oral de plus de 200 000 mots. Celle-ci regroupera différents domaines d'application intéressant actuellement la communication homme - machine (réservation hôtelière, renseignement touristique, renseignement administratif, portail vocal entreprises ou accueil téléphonique). Elle sera constituée d'ici la fin de l'année 2004.

Diffusion des ressources orales : PAROLE PUBLIQUE, ANANAS, ASILA, RTP 14 — Il n'est pas inutile de préciser que les ressources (outils et corpus) réalisées dans le cadre du projet OURAL seront distribués librement auprès des centres de recherche académiques. Cette politique recoupe totalement nos objectifs en matière de ressources linguistiques.

En effet, si notre motivation première reste la constitution de corpus pour nos propres besoins, cette activité s'accompagne d'une diffusion systématique et totalement libre de chacune de nos réalisations. Cette politique, qui est structurée par un programme interne appelé PAROLE PUBLIQUE (cf. chapitre 2, § 2.2), répond à une double motivation qui me tient particulièrement à cœur :

- D'une part, il s'agit d'amorcer un cercle vertueux en faveur d'une large diffusion des corpus francophones réalisées sur fonds publics. Outre l'évidente exigence déontologique qu'il sous-tend, cet effort est nécessaire au **développement** de l'ingénierie des langues en français.
- D'autre part, il s'agit de donner une plus grande visibilité au groupe CORAIL à travers ses réalisations en matière de ressources linguistiques orales. En particulier, l'ensemble des ressources qui seront à notre disposition à la fin de l'année 2004 devrait représenter le plus grand corpus francophone de dialogue oral *distribué librement*.

Je cherche par ailleurs à accompagner toutes les initiatives en faveur d'une plus grande utilisation de ces ressources. Le groupe CORAIL est ainsi fournisseur de ressources (corpus OTG et MASSY) pour les actions **ASILA**[28] et **ANANAS**[29] du CNRS, qui portent respectivement sur l'étude du dialogue (oral ou écrit) et sur l'annotation anaphorique de corpus.

---

[25] Valli A. et Véronis J. (1999) Etiquetage grammatical de corpus de parole : problèmes et perspectives. *Revue Française de Linguistique Appliquée*, 4(2), 113-133.

[26] Gasiglia N. (2002) Vers un corpus thématisé de dialogues radiodiffusés : défense et illustration. Actes *2ème journées de la Linguistique de Corpus*, Lorient, France. p. 19 (résumé)

[27] Le groupe CORAIL n'intervient dans ce projet qu'en qualité de fournisseur de ressource linguistique.

[28] ASILA : Action spécifique « Interaction Langagière et Apprentissage » du CNRS. Toile : http://www.loria.fr/projets/asila/

[29] ANANAS : Action « Annotation Anaphorique pour l' Analyse Sémantique de Corpus » du programme interdisplinaire « Société de l'Information » du CNRS. Toile : http://www.inalf.fr/ananas/

A la suite des activités de l'AS ASILA, je participe actuellement à une réflexion commune aux RTP 14 et 38 (département STIC du CNRS) sur la création d'une plate-forme nationale pour la diffusion de corpus de dialogue oraux ou multimodaux. Cette collaboration se poursuit actuellement par la mise en place d'une Equipe Projet Multi-Laboratoires (EPML « 50 ») intitulée « Corpus d'interaction langagiers » sous la direction de Daniel Luzatti (LIUM, Le Mans) et Anne Lacheret (CRISCO, Caen). Cette proposition, qui vise la mise en commun de ressources centrées sur l'interaction orale, s'inscrit dans un cadre résolument pluridisciplinaire. Elle réunit des centres de recherche en linguistique de l'interaction, en psycholinguistique, en dialectologie ainsi qu'en ingénierie des langues appliquée à la communication homme-machine[30]. Ces laboratoires disposent d'une expérience parfois trentenaire en matière de collecte de corpus oraux. Au cours de ces années, chaque champ disciplinaire a développé des pratiques spécifiques à ses besoins. La mise en place de la plate-forme nécessitera donc un effort conséquent de standardisation des méthodologies de recueil et de codage des ressources.

Normalisation des ressources linguistiques : RNIL — Pour être réellement profitable à l'ensemble de la communauté, le développement des ressources linguistiques doit s'accompagner d'un tel effort de normalisation. L'absence d'interopérabilité entre les corpus distribués et les systèmes les utilisant rend en effet impossible, ou tout du moins coûteuse, l'utilisation de ces ressources.

Les langages de balisage tels que XML (ou SGML antérieurement) offrent des solutions techniques pour la définition de standards d'encodage ou d'annotation. Ils ont ainsi été utilisés dans les versions successives de la TEI (*Text Encoding Initiative*), qui a constitué une étape essentielle vers la normalisation des ressources linguistiques.

Bien des efforts restent cependant à mener en la matière. A ce titre, je participe depuis décembre 2002 aux activités de la commission « Ressources Normalisées en Ingénierie de la Langue » (RNIL) de l'AFNOR. Présidée par Eric de la Clergerie (INRIA Rocquencourt), cette commission réunit différents experts francophones du domaine et est mandatée pour représenter la contribution française auprès du sous comité TC37/SC4 « *Language Resource Management* » de l'ISO dirigé par Laurent Romary (LORIA, Nancy). L'objectif de ces travaux est de produire un certain nombre de normes ou de recommandations qui serviront de référent aussi bien à la communauté scientifique qu'aux acteurs industriels du domaine[31]. Pour l'heure, les activités de la commission concernent des cadres normatifs relativement généraux tels que par exemple, des schémas de définition génériques de structures de traits.

6. CONCLUSION : UN PROGRAMME SCIENTIFIQUE POUR UN GROUPE DE RECHERCHE EMERGENT

Lorsque je fais le bilan de ces années de recherche, il me semble que je suis parvenu à un objectif qui me tenait particulièrement à cœur lors de ma nomination : celui de développer un programme de recherche en toute autonomie. Si le groupe de recherche que j'ai créé présente un caractère émergent, il n'a jamais joué un rôle de faire-valoir dans les différentes collaborations où il était impliqué. Bien au contraire, nombre d'entre elles ont été lancées sur mon initiative, dans l'espoir de donner une plus grande portée aux thématiques les plus

[30] Les laboratoires travaillant sur cette plate-forme se sont réunis en juin 2003 à l'occasion d'un atelier « Corpus » organisé à La Bresse dans le cadre de l'AS ASILA : LORIA (Nancy), ICARE (Lyon), CRISCO (Caen), LACITO (Paris VII), LIMSI (Orsay), LIUM (Le Mans), GREYC (Caen), Département des Sciences du Langage de l'Université d'Orléans et VALORIA.

[31] Rappelons qu'un des objectifs visés par l'ISO depuis sa création est de favoriser le développement de l'économie et des échanges commerciaux grâce à ses procédures de normalisation.

originales du projet de recherche qui me motive.

Bien entendu, ces réalisations ont leur limites. C'est ainsi qu'on peut regretter que le groupe CORAIL ne se soit pas investi plus fortement dans des projets européens ou plus modestement des programmes de type RNTL[32]. Avouons-le clairement : ce genre de collaboration lourde est à la limite des capacités de notre petit groupe, qui aura déjà bien à faire avec les trois projets TECHNOLANGUE auxquels il participe.

Cette taille critique insuffisante limite parfois nos ambitions en termes de réalisations. Il n'en reste pas moins que la pertinence de nos travaux me semble de mieux en mieux reconnue ces dernières années. En témoignent — peut-être ! — mes participations comme lecteur occasionnel aux revues *TAL*, *RIA*, *RIHM*, *EJOR* ou encore aux comités de programme de conférences telles que *HCP'2003*, *TALN'2003* ou *RECITAL'2004*.

Quoiqu'il en soit, je continue à défendre vaille que vaille une approche qui présente, je crois, certaines originalités par rapport aux recherches dominantes du domaine. En guise de conclusion à ce préambule, je voudrais donc rapidement situer mes activités de recherche et celles du groupe que j'anime. Ce rapide tour d'horizon sera l'occasion d'introduire les motivations scientifiques sous-jacentes à nos travaux. Celles-ci constitueront le fil rouge de ce mémoire.

6.1. Dialogue homme - machine centré sur l'utilisateur

Comme je l'ai précisé en début de chapitre, les activités du groupe CORAIL sont centrées sur la problématique du dialogue homme – machine. Plus précisément, nous nous intéressons à toute interaction langagière faisant intervenir un utilisateur humain et un système informatique. Le dialogue homme - machine rentre bien entendu dans cette définition, mais également la communication médiée par l'ordinateur, étudiée ici au titre de l'aide à la communication pour les personnes handicapées. Comme je l'ai évoqué, ces deux domaines d'application sont abordés en plaçant l'utilisateur au centre de notre réflexion.

Le rôle central de l'utilisateur dans la conception des systèmes informatiques constitue une des règles d'or des recherches en Interaction Homme - Machine. Les personnes ayant travaillé sur le monde du handicap savent ainsi que les apports d'une aide efficace à la saisie de texte peuvent être totalement ruinés par une interface mal conçue, ne pouvant pas s'adapter à chaque handicapé. Cependant, cette prise en compte des utilisateurs réels ne doit pas se limiter à une question ergonomique. Elle doit constituer la ligne directrice de nos réflexions sur les traitements linguistiques mis en jeu dans l'interaction. Pour paraphraser le titre d'un article de Donna Harman[33], nous devons passer d'interfaces centrées sur l'utilisateur à des systèmes d'ingénierie linguistique centrés sur l'utilisateur.

Cette affirmation est une évidence pour les chercheurs travaillant sur la gestion du dialogue homme - machine. Tout en s'intéressant aux aspects langagiers du dialogue, ils sont en effet en première ligne sur le front de l'échange avec l'utilisateur final[34]. Aussi ne s'étonnera-t-on pas que les recherches en pragmatique soient souvent celles qui manifestent la plus grande ouverture vers les sciences cognitives en général.

---

[32] RNTL = Réseau National des Technologies Logicielles –Toile : http://www.industrie.gouv.fr/rntl/

[33] Donna Harman (1992) User-Friendly Systems Instead of User-Friendly Front-Ends. *JASIS* 43(2): 164-174.

[34] Ce document n'abordera pas la question essentielle de la modélisation du dialogue. Comme ouvrages de référence en français, on citera : Sabah G., Vivier J., Vilnat A., Pierrel J-M., Romary L, Nicolle A. (1997) Machine, langue et dialogue ; L'Harmattan, Paris, France ; Bilange E. (1992) Dialogue personne - machine : modélisation et réalisation informatique. Hermès, Paris, France ; Pierrel J-M. (1987), Dialogue oral homme - machine. Hermès, Paris, France.

A l'opposé, ce souci me semble moins présent dans les recherches concernant les niveaux « inférieurs » de traitement. Certes, tout système de reconnaissance ou de compréhension automatique de la parole ambitionne de traiter des énoncés spontanés issus d'interactions réelles. Tout au long de ce document, je chercherai cependant à montrer en quoi les recherches actuelles ne sont pas toujours à la mesure de cet objectif. Ainsi, la communication homme - machine (CHM) et le traitement automatique des langues naturelles (TALN) ont connu au cours des deux dernières décennies une révolution paradigmatique qui s'est traduite par l'émergence d'une réelle ingénierie de la langue. Comme je le rappellerai au chapitre suivant, cette évolution a été salutaire sur bien des aspects. Il est cependant à craindre qu'elle ait également favorisé une démarche purement ingénierique qui privilégie une recherche d'efficacité à court terme au détriment d'études amont visant à comprendre le comportement langagier des utilisateurs.

Or, seule une analyse des usages langagiers me semble à même de conduire le traitement des langues naturelles vers des systèmes efficaces, robustes et surtout répondant aux besoins des utilisateurs. Car, si nous assistons actuellement à un développement très significatif des technologies langagières, on peut encore s'interroger sur le public qu'elles touchent réellement. Dans mon esprit, cette réflexion concerne en premier lieu la communication homme - machine. En effet, seuls des utilisateurs motivés ou réceptifs aux nouvelles technologies semblent prêts à utiliser des systèmes interactifs qui n'atteignent pour l'heure une certaine robustesse qu'en imposant un dialogue directif voire des contraintes d'élocution encore plus fortes. En témoignent les expériences menées avec le système néerlandais de réservation ferroviaire qui a été mis en place dans le cadre du projet européen ARISE[35]. Pour le grand public, le dialogue homme - machine reste confiné à des applications si restreintes qu'on peut s'interroger sur leur nature communicationnelle. Celles-ci se limitent en effet bien souvent à une navigation dans une arborescence figée et peu profonde. Dans ce cas, le recours à une sélection par touches DTMF semble suffire à l'obtention de l'information recherchée.

6.2. Une recherche ancrée linguistiquement

On l'aura compris, les recherches actuelles en ingénierie des langues ne me semblent pas tenir suffisamment compte du fait linguistique. Au contraire, notre groupe de recherche tente d'appréhender l'utilisateur dans sa dimension langagière.

Cette démarche m'a conduit à fonder nos systèmes de compréhension de parole sur des traitements linguistiques issus de ce que l'on appelle le « TAL robuste »[36] et non pas sur des méthodes sélectives *ad hoc*. En intégrant ces techniques d'analyse linguistique, notre objectif n'est pas de construire des systèmes anthropomorphiques. Simplement, comme je tenterai de le montrer plus loin (chapitre 4, §1 et 2), cet ancrage linguistique semble le seul à même d'atteindre un degré de généricité qui fait défaut aux systèmes de compréhension actuels. En particulier, cette voie semble autoriser à terme un dialogue homme - machine sur des domaines d'application plus riches que ceux envisagés actuellement.

Il n'en reste pas moins que les travaux du groupe CORAIL se situent toujours dans une perspective ingénierique, seule à même de conduire à des applications opérationnelles. Par exemple, la réutilisabilité des modules de traitement linguistique, ou plus simplement la cohérence des méthodes utilisées, constituent une de nos préoccupations. On constatera de

---

[35] den Os E., Boves L., Lamel L., Baggia P. (1999). Overview of the ARISE project, Actes *6th European Conference on Speech Communication and Technology*, *Eurospeech'99*, Budapest, Hongrie. 1527-1530.

[36] " *Robust parsing* en anglais ". Pour une introduction sur ces approches mettant l'accent sur la robustesse d'analyse, on pourra consulter : Carroll J., Briscoe T. (1996) Robust parsing : a brief overview. Actes *ESSLI'1996 Robust Parsing Workshop*.
Disponible sur la Toile : http://www.cogs.susx.ac.uk/lab/nlp/carroll/papers/esslli96.pdf.

même (chapitre 4 § 3) que je n'hésite pas à orienter certains travaux vers des techniques empiristes (modèles stochastiques de langages) lorsqu'une étude préalable a indiqué qu'elles étaient adaptées au problème considéré.

Dans notre approche, cette étude préalable ne peut correspondre qu'à une analyse détaillée des usages langagiers menée sur un corpus représentatif de la tâche. De fait, nos travaux en linguistique de corpus ne sauraient se limiter à la conception des systèmes interactifs. En règle générale, la conception des systèmes interactifs repose sur l'étude de corpus de dialogue permettant à la fois de circonscrire la tâche étudiée et de mieux appréhender la structure des énoncés et du dialogue que le système aura à traiter. Il s'agit avant tout de bonnes pratiques ingénieriques consistant à se confronter à la réalité que le système devra affronter. Dans ce mémoire, je tenterai de montrer, à partir d'études de cas, que la limitation des études de corpus à de tels besoins *ad hoc* nous prive d'éléments de réflexion utiles pour la conduite de nos recherches futures. Ainsi, l'ingénierie des langues à fort ancrage linguistique que j'appelle de mes vœux ne saurait se passer d'une linguistique de corpus tournée vers les besoins du traitement automatique des langues. L'orientation plus informatique pris par la linguistique de corpus au cours de ces dernières années m'apparaît de ce point de vue de bonne augure.

Le développement récent de l'ingénierie des langues s'est accompagné d'un recours croissant à des programmes d'évaluation impliquant de multiples centres de recherche. Ces campagnes de tests donnent toujours une photographie intéressante de l'état de l'art en vigueur. Reposant sur des métriques globales très générales — on les retrouve couramment en génie industriel — ces évaluations manquent cependant de pouvoir diagnostic et souffrent d'un manque de généricité. Si l'ingénierie des langues veut éviter certains aveuglements, elle devra s'appuyer sur une évaluation plus fine capable de fournir un objectif précis du comportement linguistique des systèmes qu'elle développe. C'est cet objectif que visent les méthodologies DCR et DEFI que j'ai proposé avec certains collègues. Comme je l'ai déjà évoqué brièvement, celles-ci ont inspiré pour partie la campagne d'évaluation EVALDA-MEDIA du programme TECHNOLANGUE du Ministère de la Recherche.

Ce mémoire a pour objectif de présenter mes activités de recherche à la lumière du programme scientifique qui les sous-tend : celui d'une ingénierie des langues à fort ancrage linguistique. Dans un premier temps (chapitre 1), je proposerai ma vision de l'évolution du traitement automatique des langues naturelles. L'observation des mutations qu'a connu ce domaine au cours des dernières décennies, ses réussites et ses échecs, ont en effet renforcé ma conviction sur la nécessité de combiner démarche ingénierique et recherches linguistiques. Après ce tour d'horizon critique des recherches en TAL, les chapitres qui suivront présenteront mes travaux. Les chapitres 2 et 3 seront consacrées aux recherches que je mène en amont (chapitre 2 : analyse de usages langagiers) et aval (chapitre 3 : évaluation) des activités de conception. Le chapitre 4 présentera enfin les systèmes de compréhension de parole et d'assistance aux handicapés réalisés sous ma direction. L'étude de leurs performances et de leurs comportement montrera, je l'espère, la pertinence de la démarche que je cherche à suivre.

Il est clair cependant que cet exposé ne saurait avoir valeur de démonstration absolue. La jeunesse et la taille de notre groupe de recherche limitent en effet parfois nos efforts et nos ambitions. J'espère néanmoins que l'exposé de nos travaux sera suffisamment significatif pour suggérer l'intérêt d'une ingénierie de la langue plus linguistique.

1. Pour une ingénierie des langues plus linguistique

Avoir toujours raison,
C'est un grand tort.

Egdar Faure, *Mémoires I*

1. TALN ET INTELLIGENCE ARTIFICIELLE

L'histoire des technologies langagières — qu'elles concernent le langage écrit ou le traitement de la parole — est aussi vieille que celle de l'informatique. Dès que furent disponibles les premiers calculateurs réellement programmables, nombreux ont été les pionniers de l'informatique qui se sont intéressés au problème du traitement automatique des langues (TAL par la suite). Les travaux actuels en TAL reposent donc sur les acquis de plus de cinquante années de recherches. Cependant, l'évolution du domaine fut loin d'être linéaire. Ainsi, en dépit d'indéniables avancées théoriques, mais aussi de quelques réussites techniques[1], l'histoire du TAL fut souvent celle des rendez-vous manqués et des désillusions cruelles[2].

Marqué par une forte identification entre traitements linguistiques et algorithmique, le TAL de première génération (années 1950) a plus influencé l'évolution de l'informatique qu'il a fourni des solutions efficaces au traitement automatique des langues[3].

Ce constat s'adresse également au TAL de seconde génération, qui s'est structuré à partir des années 1960 autour de la théorie des langages formels[4] et la construction d'analyseurs à base de connaissances. Dès lors, le TAL ne fit que suivre l'évolution de l'Intelligence Artificielle[5]. Il en reprenait le projet cognitiviste en recherchant de manière privilégiée la modélisation informatique des lois cognitives présidant aux fonctions langagières. La réalisation de systèmes opérationnels apparaît dès lors comme une simple conséquence de ce programme de recherche : une fois ces mécanismes cognitifs connus, la réalisation de systèmes robustes et efficaces doit dans cette démarche se limiter à un simple problème algorithmique.

On le sait, la confrontation de ces idées à la réalité fut rude et les résultats décevants[6]. Sans faire un retour historique sur les échecs de l'Intelligence Artificielle, on peut relever quelques difficultés qui s'opposèrent à la réussite de cette entreprise :

a) en dépit de nombreuses recherches en psycholinguistique, on ne dispose que d'une connaissance très parcellaire sur les mécanismes cognitifs supérieurs mis en jeu lors des activités langagières. Par exemple, un problème aussi bien délimité que les préférences de rattachement des compléments prépositionnels n'a à ma connaissance pas encore été expliqué par un modèle cognitif satisfaisant[7]. Si l'on se limite à une description purement linguistique des activités langagières, on peut de même juger que nos connaissances restent insuffisantes. Après avoir

---

1 A titre d'exemple, le système de traduction automatique SYSTRAN, opérationnel et commercialisé depuis de nombreuses années, a pour origine les recherches menées à l'université de Georgetown dans les années cinquante et soixante : Henisz - Dostert B., Macdonald R., Zarechnak M. (Eds.) (1979) Machine Translation. Mouton De Gruyter, Berlin, Allemagne. Ouvrage cité dans : Church W.K., Mercer R. L. (1993) Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1) 1-25.

2 Voir par exemple le célèbre rapport ALPAC qui a conduit en 1966 le gouvernement à réduire fortement son soutien financier aux recherches en traduction automatique : ALPAC : Automatic Language Processing Advisory National Research Council (1966) Language and machines ; computers in translation and linguistics, rapport 1416, National Academy of Sciences, Washington, Etats-Unis.

3 Remarquons tout de même que ces pionniers adoptaient une approche empiriste que ne renieraient pas les chercheurs actuels. Les critiques de Chomsky (études statistiques en linguistique), puis de Minsky et Papert (réseaux de neurones) allaient malheureusement conduire à un longue absence de confrontation aux données réelles.

4 Chomsky N. (1957) Syntactic Structures. Mouton & Co., La Haye, Pays-Bas.

5 Sabah G. (1989) L' Intelligence Artificielle et le langage. Hermès, Paris. 2 volumes.

6 Pour une vision historique et chronologique de ces temps pionniers qui ont vu la création de l'ATALA et des désillusions qui s'ensuivirent, on pourra consulter : Léon J. (1992) De la traduction automatique à la linguistique computationnelle. Contribution à une chronologie des années 1959-1965. *TAL*, 1992(1-2), 25-44.

7 Spivey - Knowlton M. J. (1992) Another context effect in sentence processing : implications for the principle of referential support, *14th Annual Conference of the Cognitive Science Society*, Bloomington, USA. 486-491.

discuté du cas des coordinations, Anne Abeillé et Philippe Blache jugent d'une manière plus globale que[8] :

« l'accumulation des connaissances syntaxiques est encore insuffisante »

b) les traitements perceptifs de bas niveau semblent difficilement répondre à une modélisation analytique des approches à base de connaissances. D'où les échecs de la reconnaissance de la parole envisagée comme un décodage acoustico-phonétique[9]. Ce constat se retrouve pour les méthodes analytiques de reconnaissance des formes utilisées alors en traitement d'images.

c) quand bien même l'élication (analyse d'expertise humaine par exemple) et la modélisation d'une telle connaissance s'avère possible, cette dernière se révèle très friable dès que la base de règles résultante atteint une taille respectable. On observe ainsi que l'ajout de règles pour prendre en compte des observations non encore modélisées induit le plus souvent des incohérences préjudiciables avec la base existante.

d) Enfin, les fonctions langagières supérieures mobilisent en parallèle des connaissances multiples (prosodie, syntaxe, sémantique, pragmatique…). La modélisation de ces interdépendances et influences complexes n'a pas encore connu de réponse cognitive satisfaisante. Elle renforce de plus la complexité des systèmes et se traduit par une importance accrue des conflits entre ces différentes sources de connaissances.

Par delà l'intérêt de parfaire nos connaissances sur la cognition humaine, les difficultés relevées ci-dessus questionnaient l'Intelligence Artificielle dans ses fondements. Certaines propositions tentèrent d'y répondre : logique floue pour une gestion plus graduelle des incohérences, ou encore Intelligence Artificielle Distribuée (systèmes multi - agents) pour la modélisation des relations entre bases de connaissances différentes[10]. Aucune ne fut en mesure de répondre complètement aux difficultés des technologies langagières. Les approches à base de connaissances semblaient ainsi devoir toujours connaître des problèmes rédhibitoires de couverture linguistique et de robustesse.

2. INGENIERIE DES LANGUES ET APPROCHES PROBABILISTES

2.1. Emergence des approches statistiques en reconnaissance de la parole

Il fallut attendre les années 1980 pour voir certains chercheurs s'engager dans une voie résolument différente reposant sur la théorie de l'information que Shannon avait développé dès la fin de la seconde guerre mondiale[11]. Cette approche mathématique, qui ne partageait aucun lien avec l'Intelligence Artificielle, fut dans un premier temps appliquée à la reconnaissance automatique de la parole. Le problème fut reformulé suivant le modèle du canal bruité élaboré par Shannon (figure 1.1). Il s'apparente à la reconnaissance d'un signal I (séquence de mots correspondant à l'énoncé prononcé) transmis à travers un canal bruité — virtuel…— pour donner un signal O observé (la séquence audio correspondant à la parole prononcée).

	I ———— canal bruité ————> O		
énoncé	*prononcé*	reconnaissance	*signal de parole*

[8] Abeillé A., Blache P. (2000) Grammaires et analyseurs syntaxiques. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 51-76 (citation p. 59-60).

[9] Méloni H. (1982), Etude et réalisation d'un système de reconnaissance automatique de la parole continue, Doctorat d'État, Université d'Aix - Marseille II, France ; Caelen J., Nasri M.K., Reynier E., Tattegrain H. (1990) Architecture et fonctionnement du système DIRA. De l'acoustique aux niveaux linguistiques. *Traitement du signal*, 7 (4). 345-366.

[10] Sabah G. (1992) Collaboration des sources de connaissances dans un système de traitement automatique des langues : l'exemple de CARAMEL, *in* J. Caelen (Ed.), *Cognition, perception et action en communication parlée* ; Roussalany A., Pierrel J.-M. (1992) Dialogue oral homme - machine en langage naturel : le projet DIAL, *Techniques et Sciences Informatiques*, 11 (2), 45:91.

[11] Shannon C. (1948) The mathematical theory of communication. *Bell System Technical Journal*, 27, 398-403.

Figure 1.1 *— Le problème de la reconnaissance de parole reformulé dans le cadre de la théorie de l'information (modèle du canal bruité).*

La théorie de l'information indique que le problème de la reconnaissance du signal I à partir de l'observation O revient à rechercher la séquence I qui, parmi toutes celles envisageables, maximise la probabilité d'avoir I connaissant l'observation O :

$$I_{solution} = \underset{I}{\text{argmax}}\ Pr(I \mid O)$$

En pratique, il est plus facile d'estimer cette probabilité sous la forme suivante, obtenue par application de la règle de Bayes :

$$I_{solution} = \underset{I}{\text{argmax}}\ Pr(I) \,.\, Pr(O|I)$$

Pr(I) est la probabilité d'émission de la séquence de mots (W_{I1}, W_{I2}, … , W_{In}) représentant l'énoncé I. Elle modélise implicitement le langage étudié, puisque la probabilité d'observer un énoncé ne dépend que de contraintes lexicales[12], syntaxiques[13] et sémantico-pragmatiques[14]. Aussi parle-t-on généralement de modèles statistiques de langage.

De son côté, Pr(O|I) est la probabilité d'observation du signal de parole (séquence acoustique) O compte tenu de l'énoncé I prononcé. Elle rend compte des niveaux infra - linguistiques mis en jeu par la production orale. On parle ici généralement de modèles acoustiques.

La complexité du modèle statistique ainsi formulé dépasse de loin les capacités de calcul et d'apprentissage des systèmes informatiques. Du fait de l'existence de dépendances à longues distances ou des phénomènes de co-référence dans les langues naturelles, la probabilité d'apparition d'un mot devrait en effet dépendre de l'ensemble des mots déjà prononcés par le locuteur ! Il est donc nécessaire de simplifier l'estimation de ces probabilités à l'aide d'hypothèses réductrices qui restent acceptables au regard du problème posé. Dans le cas des modèles de langage statistiques, on postule ainsi que la probabilité d'occurrence d'un mot ne dépend que de ses N-1 prédécesseurs (*N-grammaires*), avec des valeurs de N inférieures ou égales à … trois.

Au final, le problème est décrit par un modèle paramétrique formalisé par des chaînes de Markov cachées (*HMM* pour *Hidden Markov Models* en anglais). Les paramètres du modèle sont estimés automatiquement sur de grands corpus. L'obtention de la solution se ramène à un problème de recherche dans un espace de probabilités. On notera que l'on disposait dès le début des années 1970 de solutions algorithmiques efficaces pour l'estimation paramétrique et la recherche de solutions de probabilité maximale[15].

2.2. Une révolution empiriste progressivement acceptée par le TAL

Cette approche fondée sur une modélisation purement probabiliste retournait complètement le paradigme dominant en vigueur : à la démarche centrée sur la connaissance de l'Intelligence Artificielle succédait une approche centrée sur les données (apprentissage sur corpus). Pour saisir l'importance de ce changement épistémologique, il faut insister sur le fait que la finalité de ces

[12] Par exemple, l'observation d'un mot fréquent dans la langue est plus probable que celle d'un mot rare ou inconnu.

[13] Le caractère agrammatical de l'énoncé " *le chapeau beau de mon père* " rend son observation moins probable que la séquence normative qui lui correspond (" *le beau chapeau de mon père* "). Remarquons que la notion d'acceptabilité n'a plus lieu d'être en théorie de l'information. L'apparition de l'énoncé agrammatical n'est en effet pas interdite. Simplement, son observation étant rare, sa probabilité d'occurrence sera faible.

[14] Le célèbre exemple de Noam Chomsky " *colourless green ideas sleep furiously* ", bien que régulier d'un point de vue syntaxique, n'a que peu de chance d'être observé du fait de son incohérence. Cette plausibilité sémantique prend une importance cruciale dans le cadre d'applications finalisées. En effet, un énoncé jugé parfaitement acceptable dans un contexte (pragmatique) donné peut apparaître comme totalement dénué de sens dans un autre.

[15] Par exemple, pour l'estimation des modèles markoviens : Baum L.E., Eagon J.A. (1967) An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*. 73, 360-363.

modèles statistiques n'est pas la découverte d'une propriété linguistique sous-jacente mais bien la recherche d'un comportement optimal du modèle, quand bien même celui-ci n'est pas explicable. De ce point de vue, ces approches ne partagent que peu d'affinités avec les techniques d'apprentissage relevant de l'inférence grammaticale. A l'opposé, on peut les rapprocher des méthodes connexionnistes[16].

Or, ce changement de paradigme ne concerne pas que la reconnaissance de la parole. Au contraire, il est susceptible de s'appliquer de nombreuses problématiques relevant du traitement du langage naturel. C'est ce que suggèrent les figures 1.2 et 1.3, qui reformulent les problématiques de la traduction automatique et l'annotation morpho - syntaxique dans le cadre de la théorie de l'information.

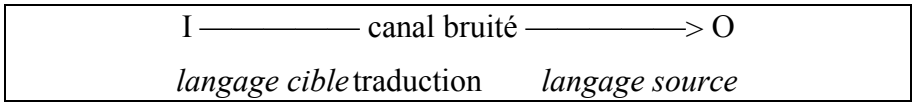

Figure 1.2 *— Le problème de la traduction automatique reformulé dans le cadre de la théorie de l'information (modèle du canal bruité).*

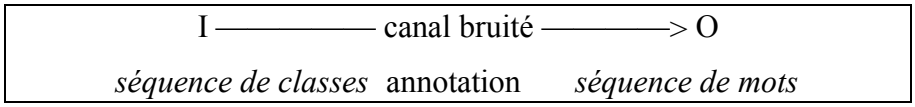

Figure 1.3 *— Le problème de l'annotation morpho - syntaxique (assignation de classes syntaxiques) reformulé dans le cadre de la théorie de l'information (modèle du canal bruité).*

Ce changement de cadre conceptuel n'allait pas de soi pour une communauté scientifique qui avait développé depuis près de 25 ans une approche rationaliste et analytique. A la suite de travaux pionniers menés en dictée vocale, cet aggiornamento fut cependant rapidement accepté par la communauté du traitement de la parole. Une des explications à cette adhésion réside certainement dans la proximité du traitement du signal, largement utilisé dans le domaine, avec la théorie de l'information.

A l'opposé, l'irruption de méthodes stochastiques fut ressentie plus douloureusement dans le domaine du langage écrit[17]. D'où des réticences qui étaient encore sensibles il y a peu, en particulier au sein du TALN francophone. Ainsi, Marc Dymetman pouvait-il écrire au milieu des années 1990 au sujet de la traduction automatique[18] :

> *" Bien que l'on puisse à bon droit rester sceptique sur la capacité de ces méthodes à produire des traductions acceptables, il faut reconnaître que certains résultats obtenus [...] ont surpris les spécialistes du domaine, non pas peut-être par leur qualité,* ***mais du simple fait de leur existence*** *"* (souligné par nous).

Certains résultats prometteurs obtenus par les approches statistiques, auxquels répondaient les limitations constatées des méthodes à base de connaissances, levèrent cependant progressivement

---

16 Il a d'ailleurs été montré qu'un modèle de Markov caché (HMM) peut être simulé par un réseau de neurones récurrent entraîné avec l'algorithme de rétro - propagation dans le temps : Bridle J.S. (1990) Alpha-nets : a recurrent " neural " network architecture with a Hidden Markov Model interpretation. *Speech Communication*, 9(1), 83-92.

17 C'est essentiellement à partir des années 1990 que fut envisagée l'utilisation d'approches statistiques pour le traitement du langage écrit. Un des exemples les plus frappants de cette évolution concerna les travaux pionniers d'IBM dans le domaine emblématique de la traduction automatique. Brown P., Cocke J., Pietra S. D., Jelinek F., Lafferty J.D., Mercer R.L. et Rossin P.S. (1990) A statistical approach to machine translation. *Computational Linguistics*, 16(2), 79-85.
On citera également les travaux pionniers de Church en analyse syntaxique partielle : Church K. (1988) A stochastic parts program and noun phrase parser for unrestricted text. Actes 2nd *Conference on Applied Natural Language Processing*, *ANLP'1988*, Austin, Etats-Unis, 136-143.

18 Dymetman M. (1994) Quelques développements récents à la périphérie de la Traduction Automatique, actes *TALN'94*, Marseille, France, 24-30.

ces défiances : les approches statistiques semblaient bien avoir un rôle à jouer dans les technologies langagières.

C'est en reconnaissance de la parole que cet apport fut le premier reconnu. Ainsi, Alex Waibel pouvait affirmer dès 1990 dans l'introduction d'un ouvrage consacré au domaine[19] :

> *" The pure knowledge-based approach emulates human speech knowledge using expert-systems. Rule-based systems have had only limited success* […]. *Most successful large-scale systems today use a stochastic approach "*

Ce mouvement s'est poursuivi depuis, pour toucher finalement une part très significative des problématiques relevant du traitement du langage. En témoigne le tableau 1.1, qui présente la période d'apparition de travaux statistiques sur une sélection de domaines de recherche.

Tableau 1.1 *— Période d'apparition des travaux de nature statistiques par thématique de recherche. Chronologie estimée à partir de publications relevées dans la littérature (les publications listées dans ce tableau sont recensées dans la bibliographie donnée en fin d'ouvrage).*

Thématique	Année d'apparition	Auteurs de la publication
Recherche d'information textuelle	(1972)	(Salton)[20]
Reconnaissance automatique de la parole	1975 1976	Baker Jelinek
Etiquetage morpho syntaxique	1976 1977 1983	Bahl et Mercer Debili Leech, Garside, Atwell , Marshall
Segmentation (*shallow parsing*)	1988	Church
Traduction automatique	1990	Brown, Cocke *et al*.
Compréhension de la parole	1992	Levin et Pieraccini

La généralisation de ces méthodes novatrices a profondément marqué le traitement automatique des langues naturelles. Par delà ce changement de techniques, elle a entraîné une modification très sensible de perspective pour les recherches menées en TALN. C'est en effet avec le TAL probabiliste que naquit véritablement l'ingénierie des langues.

2.3. L'ingénierie des langues, ou le véritable apport des méthodes statistiques

Nous avons eu beau jeu de mettre en exergue jusqu'ici les difficultés auxquelles se heurte l'Intelligence Artificielle symbolique. C'est en effet oublier les limitations que présentent les modèles statistiques de langage de leur côté. Les hypothèses réductrices sur lesquels est basée l'estimation de leurs paramètres limitent en effet la connaissance linguistique qu'ils peuvent encoder. On connaît par exemple les difficultés que rencontrent les N-grammaires à modéliser les

[19] Waibel A., Lee K. (Eds) (1990) Readings in speech recognition. Morgan Kaufman, Princeton, NJ : p. 4.

[20] Les approches statistiques utilisées en recherche d'information ne relèvent pas du modèle du canal bruité (*cf.* § 4.1).

dépendances à longue distance[21]. La reconnaissance de ces insuffisances fut relativement précoce et a donné lieu à de multiples tentatives d'amélioration de ces modèles[22].

Ainsi, si les approches à base de connaissances ont souvent péché par manque de robustesse, les méthodes statistiques manquent de couverture linguistique. Comment expliquer alors qu'elles aient obtenu des résultats là où les approches symboliques avaient échoué ? A mon sens, cette explication se trouve avant tout dans l'état d'esprit qui a animé les pionniers du TAL statistique.

Comme nous l'avons vu précédemment, le TAL symbolique a longtemps suivi le programme cognitiviste de l'Intelligence Artificielle. Si la réalisation d'applications opérationnelles n'était pas absente des préoccupations, elle n'était pas au centre de la réflexion théorique des chercheurs. De même que la linguistique Chomskyenne travaillait sur des exemples créés de toutes pièces, le TAL se préoccupait d'énoncés artificiels (cas jouets) portant sur des phénomènes linguistiques aussi complexes que peu fréquents.

A l'opposé, les recherches en TAL statistique ont toujours eu pour finalité la recherche d'une robustesse optimale sur des données ou des applications réelles. Elles se donnent pour objectif prioritaire la réalisation de systèmes efficaces et opérationnels dont les performances sont estimées au cours de campagnes d'évaluation significatives portant sur des situations proches du réel.

Cette recherche d'efficacité, ce recours systématique à l'évaluation relèvent clairement d'une perspective ingénierique. Le fait que cette démarche alors novatrice se soit désormais imposée à l'ensemble du traitement automatique des langues constitue certainement un apport majeur des approches probabilistes.

2.4. L'ingénierie des langues : une démarche générale

Une dizaine d'années après la généralisation du TAL probabiliste, les recherches en traitement automatique des langues sont ainsi menées dans une perspective ingénierique. On relève par exemple une utilisation de plus en plus fréquente des termes d'ingénierie de la langue[23] ou de la parole[24]. De même, les principales revues scientifiques du domaine[25] accordent désormais une place privilégiée à ce type de travaux.

Parallèlement, on assiste à un renforcement marqué des liens entre industries de la langue et recherche scientifique. La création de manifestations telles que LangTech[26], ou encore la diffusion d'une lettre électronique sur les technologies langagières (Euromap), témoignent autant de l'essor de ce domaine d'activité que de l'intérêt des chercheurs pour les transferts de technologies. Le

[21] Cette affirmation demande toutefois à être relativisée. Plusieurs expériences sur les permugrammes (combinaison linéaires de N-grammaires agissant sur des permutations de l'historique de prédiction), multigrammes ou autres modèles markoviens dynamiques ont montré qu'il est possible d'intercepter certaines dépendances à longues distances sans recourir à un accroissement de la taille N de la fenêtre de contexte. Ces méthodes requièrent cependant un apprentissage très lourd. Pour une revue de détail : Zitouni I. (2000) Modélisation du langage pour les systèmes de reconnaissance de la parole destinés aux grands vocabulaires. Doctorat de l'Université Nancy I, France ; Deligne S., Bimbot F. (1995) Language modelling by variable length sequences : theoretical formulation and evaluation of multigrams. Actes IEEE *International Conference on Acoustics, Speech and Signal Processing*, *ICASSP'1995*, Detroit, MI. 172-175.

[22] Voir par exemple le tutoriel donné en 1991 par Frederik Jelinek sur les efforts déployés par la communauté scientifique pour améliorer les modèles de langage de type trigrams : Jelinek F. (1991) Up from trigrams ! The struggle for improved language models. *Actes 2nd European Conference on Speech Communication and Technology*, *Eurospeech'1991*, Gènes, Italie. 1037-1040.

[23] Pierrel J-M. (dir.) (2000) Ingénierie des langues. *Collection I²C*. Hermès. Paris

[24] Dans leur ouvrage consacré au traitement de la parole, Boite *et al*. s'adressent explicitement à un lectorat d'ingénieurs : Boite R., Bourlard H., Dutoit H., Hancq J. et Leich H. (2000) Traitement de la parole, Coll. Electricité, Presses Polytechniques et Universitaires Romandes, Lausanne, Suisse.

[25] On pense en premier lieu à *Natural Language Engineering*, qui a été créée explicitement dans cette perspective. De leur côté, les revues plus anciennes accordent elles aussi une place essentielle aux travaux empiristes et à l'évaluation. L'évolution éditoriale de *Computational Linguistics* peut ainsi être datée de l'année 1993, qui vit la publication d'un numéro spécial consacré à l'utilisation des corpus en TALN : Special issue on using large corpora :I. *Computational Linguistics*, 19(1), mars 1993.

[26] Le premier salon LangTech s'est déroulé à Berlin en septembre 2002 : http://www.lang-tech.org/

récent programme *Technolangue* lancé par le Ministère de la Recherche devrait être révélateur de ce renforcement des échanges.

Les technologies langagières ont par ailleurs atteint le stade des applications et des services opérationnels à destination du grand public ou des professionnels. Parmi les domaines d'application les plus en pointe, on citera la dictée vocale, la correction orthographique, l'indexation et recherche d'information textuelle, voire de manière plus anecdotique la traduction assistée par ordinateur et la mise en œuvre de serveurs vocaux interactifs (dialogue oral homme - machine). Bien que sévèrement touchées par la crise boursière actuelle, de multiples PME-PMI francophones ont une activité centrée sur les technologies langagières[27]. Par ailleurs, l'intérêt de nombreuses multinationales pour l'ingénierie des langues témoigne des perspectives de développement de ce domaine d'activité. Assez naturellement, ces grandes sociétés relèvent des domaines de l'informatique (Microsoft, IBM), des télécommunications (Telisma, essaimage de France Telecom R&D) ou du monde de l'information et des médias (Xerox, Philips, Thales).

Or, il est intéressant de noter que ces avancées reposent indifféremment sur des traitements symboliques ou probabilistes. Expression du caractère désormais global de l'ingénierie des langues, la définition qu'en donne Cunningham s'appuie uniquement sur ses finalités et non sur les approches utilisées. Il distingue ainsi[28] :

- la linguistique computationnelle (*Computational Linguistics*), qu'il définit comme un champ de la linguistique utilisant l'outil informatique à des fins de validation des théories,
- le TALN (*Natural Language Processing*) qui serait le champ de l'informatique s'intéressant au traitement de la langue, d'un point de vue pouvant être théorique.
- enfin, l'ingénierie des langues (*Language Engineering*) qui est une sorte de « TALN appliqué » reposant sur des contraintes de mise en œuvre et d'évaluation d'applications réelles. L'ingénierie des langues serait ainsi au TALN ce que le génie logiciel est à l'informatique théorique.

Cette tentative de typologie peut être discutée. Il n'en reste pas moins que l'ingénierie des langues ne peut plus être identifiée aux seules approches statistiques. Très tôt, les premiers programmes d'évaluation[29] qui marquèrent l'émergence de l'ingénierie des langues virent ainsi la participation de systèmes à base de connaissances aussi bien que probabilistes.

Un seul exemple, emblématique, suffira à démontrer l'évolution du TAL symbolique en direction de l'ingénierie des langues. L'analyse syntaxique (*parsing*) a toujours été considérée comme un champ d'investigation privilégié du TALN. Concentrant une grande partie des efforts de la communauté scientifique, il a donné lieu à de multiples propositions de formalismes[30] dont une des caractéristiques est d'être le plus souvent équivalents d'un point de vue formel[31].

---

[27] Citons par exemple les sociétés Elan, Sinequa, Lingway, Tagmatica ou encore Synapse Développement. Voir le site de l'APIL (Association des Professionnels des Industries de la Langue) : www.apil.asso.fr/entreprises_alpha.htm

[28] Cunningham H. (2000) A definition and short history of language engineering. *Natural Language Engineering*. 5(1), 1-16.

[29] Pallet D., Fiscus J., Fisher W., Garofolo J., Lund B., Prysboski M. (1994) 1993 benchmark tests for the ARPA spoken language program. Actes 1994 *ARPA Human Language Technology workshop*. Morgan Kaufman, Princeton, NJ. 49-74.

[30] LFG, HPSG, TAG, grammaires de dépendances ou catégorielles, etc. Pour une revue d'ensemble, on consultera : Abeillé A. (1993) Les nouvelles syntaxes : grammaires d'unification et analyse du français, Armand Colin, Paris.

[31] Il s'agit généralement d'équivalence faible, c'est-à-dire que les grammaires génèrent des langages identiques mais pas les mêmes arbres de dérivation. Il a été montré qu'un ensemble de grammaires parmi lesquelles les grammaires d'arbres adjoints (TAG), les grammaires catégorielles combinatoires (CCG) et les grammaires gouvernées par la tête (HG), sont faiblement équivalentes. Ce groupe de grammaires a été qualifié de faiblement sensibles au contexte (*mildly context-sensitive grammars*) : Joshi A., Weir D. et Vijay-Shanker K. (1991) The convergence of mildly context-sensitive formalisms, In Sells P., Shieber S., Wasow T. (Eds.) *Foundations issues in Natural Language Processing*, MIT Press, Cambridge, MA.

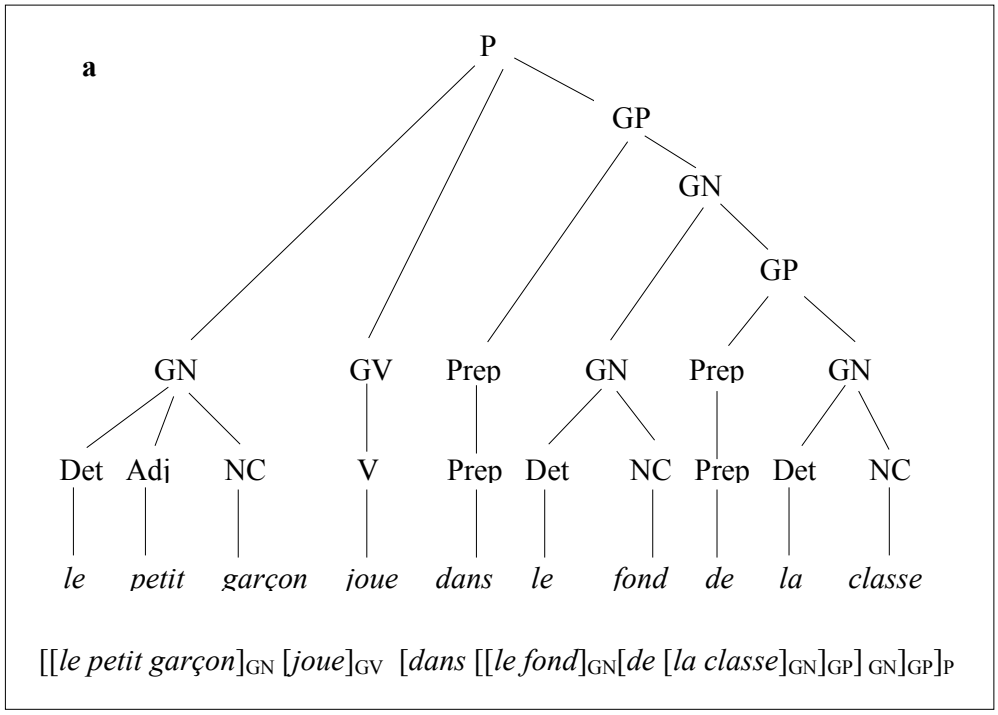

Figure 1.4 *— Analyse syntaxique complète (parsing) et segmentation en constituants noyaux (shallow parsing) de l'énoncé « le petit garçon joue dans le fond de la classe ». a) un arbre syntaxique simplifié correspondant à une parsing possible de l'énoncé b) une segmentation possible en segments minimaux de l'énoncé*

Reposant sur des motivations théoriques plus que sur des considérations pratiques, ces différentes propositions visaient une analyse syntaxique complète d'énoncés de langue générale. Moins ambitieux a priori, les premiers travaux probabilistes en analyse syntaxique[32] ont été d'emblée plus modestes. Ils se contentaient en effet de chercher à segmenter l'énoncé en constituants noyaux non récursifs (figure 1.4).

Une dizaine d'années plus tard, l'importance de ces traitements superficiels (*shallow parsing*) est reconnue de tous. Ils sont ainsi fréquemment considérés comme un premier niveau de traitement nécessaire à la mise en œuvre d'analyseurs syntaxiques complets. Nous verrons également (*cf.* chap. 4 § 1.5) que ces traitements de surface constituent une solution robuste pour la mise en œuvre de systèmes de compréhension de parole spontanée à forte composante linguistique.

Or, les approches à base de connaissances se sont elles aussi emparées de cette problématique. Dans un article alors novateur, Eva Ejerhed parle ainsi de « nouveaux courants » en analyse syntaxique[33]. Aussi les approches symboliques vont-elles représenter rapidement une part importante des recherches du domaine. On citera par exemple l'analyseur déterministe *Fidditch* développé par Hindle[34] à la même époque que les travaux fondateurs de Church sur la segmentation probabiliste.

[32] Church K. (1988) A stochastic parts program and noun phrase parser for unrestricted text. 2nd *Conference on Applied Natural Language Processing*, *ANLP'1988*, Austin, Etats-Unis, 136-143.

[33] Dans cet article, Ejerhed mettait en fait sur un pied d'égalité approches symboliques et probabilistes du *shallow parsing*. Elle même proposait en fin d'article une extension aux propositions non récursives du segmenteur probabiliste de Church : Ejerhed E. (1993) Nouveaux courants en analyse syntaxique, *TAL*, 34(1), 61-82.

[34] Hindle D. (1983) Deterministic parsing of syntactic non fluencies. Actes *21th Annual meeting of the Association for Computational Linguistics, ACL'1983*, MIT, Cambridge MS. 123-128 ; Hindle D (1989) Acquiring disambiguation rules from text. Actes *27th Annual meeting of the Association for Computational Linguistic ACL'1989*, Vancouver, Canada. 118-125.

Ou encore, dans l'espace francophone, les travaux de Jean-Pierre Chanod reposant sur une analyse déterministe à base de règles de décisions[35]. Comme de nombreux autres segmenteurs, les règles de cet analyseur étaient représentées sous forme d'automates à états finis.

3. QUELLES METHODES POUR L'INGENIERIE DES LANGUES ?

Comme le montre l'exemple du *shallow parsing*, l'irruption des techniques probabilistes a entraîné la définition de nouvelles problématiques pour le TALN. Si l'ingénierie des langues a désormais recours à des techniques aussi bien symboliques que probabilistes, on peut se demander si certaines problématiques ne sont pas plus adaptées à une approche particulière. Le bref état de l'art que nous allons brosser ci-dessous suggère qu'il n'en est rien. Au contraire, il apparaît que la palette d'intervention de chaque approche est remarquablement étendue.

3.1. Reconnaissance des formes : dictée vocale et reconnaissance de caractères

A ma connaissance, il n'est qu'un seul domaine relevant des technologies langagières pour lequel la prédominance des approches stochastiques ne souffre d'aucune contestation. Il s'agit de la reconnaissance automatique de la parole. Les systèmes de dictée vocale actuels reposent en effet sur une modélisation purement statistique[36] sous forme de chaînes de Markov cachées. Les modèles de langage utilisés sont des bigrams, des trigrams, ou plus souvent leur combinaison dans une stratégie de recherche multi-passes.

Cette prédominance des méthodes probabilistes semble se retrouver en reconnaissance optique de caractères (OCR : *Optical Character Recognition* en anglais). Prise isolément, cette problématique ne nécessite pas de prise en compte du contexte linguistique. Il s'agit d'un pur problème de reconnaissance des formes reposant sur une classification des observations. Dans un bref texte de synthèse sur le sujet[37], S. Srihari et R. Srihari mentionnent ainsi l'utilisation de réseaux de neurones artificiels ou de classificateurs statistiques classiques[38].

Cette reconnaissance hors-contexte est appropriée, par exemple, dans le cas de l'identification des codes postaux. Cependant, lorsque le problème porte sur du texte écrit ou typographié, il est intéressant de guider la reconnaissance par une analyse du contexte linguistique courant. On se retrouve ici dans une problématique très proche du couplage modèle acoustique / modèle de langage de la reconnaissance automatique de la parole (figure 1.5). L'utilisation de modèles de langage stochastiques (N-grammaires de mots ou de classes[39]) dans ce domaine n'a donc rien d'étonnant.

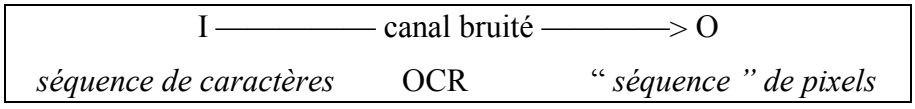

Figure 1.5 *— Le problème de la reconnaissance optique de caractères (OCR : Optical Character Recognition) reformulé dans le cadre de la théorie de l'information (modèle du canal bruité).*

[35] Chanod J.P. (1994) Développements en analyse syntaxique automatique. Actes TALN'1994, Marseille. France. 87-91. Pour une extension récente de ces travaux vers une analyse plus profonde, on consultera : Ait-Mokhtar S., Chanod J.-P., Roux C. (2002) Robustness beyond shallowness : incremental deep parsing. *Natural Language Engineering*, 8(2-3). 121-144.

[36] Nous ne discuterons pas ici de l'apport des techniques connexionnistes, en particulier au niveau de la modélisation acoustique, dans les systèmes hybrides HMM / réseaux de neurones : Bourlard H., Morgan N. (1994) Connectionist speech recognition : a hybrid approach. Kluwer Academic Publ., Dordrecht, Pays-Bas.

[37] Srihari S.N., Srihari R.K. (1995) Written Language Input. *In* Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon. Disponible sur la Toile : http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. 71-76

[38] Il est plus difficile de situer les méthodes évolutives (algorithmes génétiques), qui présentent un caractère hybride entre représentation symbolique et approche empirique utilisant une fonction d'adaptation numérique. Ces approches ont également donné des résultats significatifs en reconnaissance de caractère : Ménier G. (1995) Système en ligne de lecture d'écriture cursive manuscrite. Analyse continue des primitives et interprétation globale optimisée par algorithme génétique. Doctorat Université Rennes 1, Rennes, France.

[39] Srihari R., Baltus C. M. (1993) Incorporating syntactic constraints in recognizing handwritten sentences. Actes *13th International Joint Conference on Artificial Intelligence*, *IJCAI'1993*, Chambéry, France : 1262.

Au vu de ces deux exemples, il semble donc que les approches symboliques n'aient ainsi qu'un rôle marginal à jouer sur des tâches qui relèvent principalement de traitements perceptifs chez l'être humain. On retrouve ici les difficultés de l'Intelligence Artificielle à rendre compte de niveaux cognitifs inférieurs où une modélisation probabiliste ou connexionniste semble plus adaptée. A l'opposé, dès que l'on s'intéresse à des traitements relevant de fonctions langagières supérieures, l'utilisation de techniques à base de connaissances redevient pertinente.

Cette affirmation s'impose déjà lorsque l'on considère les modèles stochastiques de langage utilisés en reconnaissance de la parole. Nous avons signalé plus haut que ces modèles (N-grammaires de mots ou de classes) ne peuvent rendre compte complètement des dépendances à longue distance. Il ne s'agit en aucune manière d'une limitation théorique, comme le montrent les erreurs triviales que commettent les systèmes de dictée vocale. On relève ainsi de nombreuses fautes d'accord, en particulier dans le cas d'homophonies singulier / pluriel sur un même lemme. En ne considérant qu'un contexte linguistique très restreint, voire en ignorant la structure syntaxique de l'énoncé, les N-grammaires ne peuvent modéliser qu'imparfaitement ces règles d'accord relativement simples.

Deux stratégies ont été proposées pour répondre à cette limitation. La première se limite à un couplage séquentiel entre modèle stochastique et information linguistique, suivant une stratégie de recherche multi-passes[40]. L'idée est de faire une première recherche de solutions à l'aide d'une N-grammaire, puis de filtrer ou réordonner ces hypothèses à l'aide de modèles linguistiques. Cette seconde étape peut travailler indifféremment sur les meilleures séquences de mots (algorithme *N-best*[41]) ou même sur le treillis de mots[42] fournis en sortie de la reconnaissance probabiliste. Différents modèles linguistiques peuvent être utilisés dans cette étape de ré-estimation. En particulier, une grammaire symbolique peut filtrer les hypothèses jugées agrammaticales. Cette stratégie est cependant sous-optimale. Elle ne remet en effet pas en cause le classement relatif des hypothèses conservées, classement qui a été établi par un modèle stochastique dont on connaît les limitations. Il est donc préférable de ré-estimer ces hypothèses à l'aide d'un modèle probabiliste ancré linguistiquement. Il a ainsi été proposé d'utiliser des grammaire hors-contexte probabilistes[43].

Les stratégies multi-passes se retrouvent dans la plupart des systèmes de dictée vocale. On constate cependant que ces systèmes se contentent le plus souvent d'utiliser un modèle de langage d'ordre supérieur lors de la ré-estimation. Ce sont donc les avantages computationnels du couplage séquentiel qui ont retenu l'attention des ingénieurs : une première passe d'analyse, nécessairement lourde, est mise en œuvre à l'aide d'un modèle de langage relativement léger (bi-gram par exemple), puis l'utilisation de modèles d'ordre supérieur (tri-gram) permet un raffinement de la solution sur un espace de recherche limité. Tout en étant particulièrement efficace, cette stratégie ne répond pas aux déficiences des modèles de langage markoviens.

La seconde stratégie est à l'opposé beaucoup plus ambitieuse. Elle vise directement l'intégration d'informations linguistiques supérieures dans le modèle de langage initial. Certaines propositions reposent sur une hybridation entre méthodes symboliques et probabilistes, tandis que d'autres se limitent à l'intégration d'informations linguistiques (syntagmes par exemple) dans des modèles purement stochastiques. Sans être exhaustif, citons quelques travaux qui délimitent en creux les insuffisances des modèles de langage probabilistes actuels :

---

40 Schwartz R. , Nguyen L., Makhoul J. (1996) Multiple-pass search strategies. *In* Lee C.H., Soong F.K., Paliwal K.K. (Eds.) *Automatic speech and speaker recognition*. Kluwer Academic Publ., Dordrecht, Pays-Bas.

41 Schwartz R., Chow Y.-L. (1990) The N-best algorithm : an efficient and exact procedure for finding the N most likely sentence hypotheses. Actes *IEEE International Conference on Acoustics, Speech and Signal Processing*, I*CASSP'1990*., Albuquerque, NM. 81-84.

42 Su. K-Y. et al. (1992) A unified framework to incorporate speech and language information in spoken language processing. Actes *IEEE International Conference on Acoustics, Speech and Signal Processing*, *ICASPP'1992*, San Francisco, CA. vol. 1, 185-188.

43 Chappelier J.-C., Rajman M., Aragües R., Rozenknop A. (1999) Lattice parsing for speech recognition. Actes *TALN'1999*, Cargèse, France. 95-104.

- utilisation de grammaires hors-contexte probabilisées[44], voire de grammaires stochastiques à substitution d'arbres[45], pour rendre compte explicitement de contraintes syntaxiques relevant de dépendances à longue distance,
- combinaison linéaire de N-grammaires avec une grammaire tri-gram portant sur des syntagmes (et non plus sur des mots ou des classes) pour étendre la fenêtre de contexte du modèle de langage[46]. L'apprentissage de ce modèle « tri-syntagme » s'effectue sur un corpus segmenté,
- « permugramme » obtenu par combinaison linéaire de bi-grams et tri-grams agissant non pas sur les derniers mots prononcés mais sur des permutations de l'historique de prédiction. Ce modèle complexe permet d'intercepter certaines dépendances à longue distance[47]. Les « multigrammes » définis par Bimbot et Deligne poursuivent le même objectif et nécessitent également un apprentissage très lourd[48].
- modèles de langage stochastique intégrant des informations sur la structure syntaxique de surface de l'énoncé[49]. Les probabilités du modèle structurel de Chelba et Jelinek ne concernent plus les derniers mots prononcés, mais les têtes lexicales des deux derniers constituants non récursifs déjà analysés. Ces chunks sont caractérisés par un segmenteur à base d'automate à états finis probabilisés. Les auteurs montrent qu'avec le même nombre de paramètres à estimer, la fenêtre de contexte passe de 2 mots pour un trigram à 3,4 mots en moyenne pour leur modèle. Cette proposition, qui intègre partiellement la structure syntaxique de l'énoncé, présente un intérêt qui va bien au delà de la recherche d'une modélisation à contexte étendu. Nous y reviendrons.

Ces travaux, qui datent souvent de la première moitié des années 1990, dénotent une prise de conscience très précoce des limites des modèles markoviens. Les expérimentations menées sur ces modèles de langage « étendus » concluent à des améliorations sensibles en terme de perplexité[50]. Elles ne se sont pourtant pas traduites par des applications pratiques en dictée vocale. En particulier, les modèles hybrides à base de grammaires hors-contexte probabilisées restent cantonnés à des utilisations marginales telles que la contrainte de N-grammaires générales sur des sous-domaines pour lesquels on ne dispose pas de corpus d'entraînement[51].

---

44 Rajman M., Han J. (1995) Prise en compte de contraintes syntaxiques dans le cadre d'un système de reconnaissance de la parole, Actes *TALN'1995*, Marseille, France. 97-106.
D'autres auteurs proposent de combiner une grammaire hors-contexte probabilisée à des modèles stochastiques classiques (tri-gram) : Gillet J., Ward W. (1998) A language model combining trigrams and stochastic context-free grammars. Actes *5th International Conference on Spoken Language Processing*, *ICSLP'1998*, Sidney, Australie. 2319-2322.

45 Rozenknop A., Chappelier J.-C., Rajman M. (2003) Apprentissage discriminant pour les grammaires à substitution d'arbres. actes TALN'2003, Batz-sur-Mer, France. 225-234.

46 Ici, l'étude proposée visait uniquement la résolution des homophonies singulier / pluriel en dictée vocale. La portée de ce modèle est cependant plus générale : Béchet F., Nasr A., Spriet T., De Mori R. (1999) Modèles de langage à portée variable : application au traitement des homophones. Actes *TALN'1999*, Cargèse, France, 35-44.

47 Schukat-Talamazzini E.G., Hendrych R., Kompe R., Niemann H. (1995) Permugram language models, Actes *4th European Conference on Speech Communication and Technology*, *Eurospeech'95*, Madrid, Espagne, 1773-1776.

48 Deligne S., Bimbot F. (1995) Language modeling by variable length sequences : theoretical formulation and evaluation of multigrams. Actes *International Conference on Acoustics, Speech and Signal Processing*, *ICASSP'1995*, Detroit, MI, 172-175.

49 Chelba C., Jelinek F. (2000) Structured language modeling. *Computer Speech and Language*, 14(4), 283-332.
Dans le même numéro de la revue, un autre article propose la combinaison de ce modèle structurel avec une spécialisation sur le thème de l'énoncé (unigram dépendant du domaine) : Khudanpur S., Wu J. Maximum-entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling. *Computer Speech and Language*, 14(4), 283-332.

50 Notion issue de la théorie de l'information, la perplexité est une mesure d'entropie qui estime la complexité conjointe d'une tâche donnée — ici le langage à traiter — et du modèle statistique — ici le modèle de langage — utilisé pour la décrire. A tâche égale, une réduction de la perplexité traduit une réduction de la complexité du modèle, donc *a priori* une modélisation plus efficace. Nous discuterons ultérieurement (chapitre 3 § 1.2) des limitations de ce type de métriques purement ingénieriques. Remarquons simplement ici qu'il est possible de faire une analogie grossière entre la perplexité et le facteur de branchement utilisé en TAL symbolique.

51 Wang Y., Mahakan M., Huang X. (2000) A unified context-free grammar and N-gram model for spoken language processing. Actes *International Conference on Acoustics, Speech and Signal Processing*, *ICASSP'2000*, Istambul, Turquie, 1639-1642. Communication citée dans : Huang X., Acero A., Hon H-W. (2001) Spoken language

Cette situation paradoxale nous questionne sur l'apport des modèles stochastiques de langage en reconnaissance de la parole. Est-ce leur intégration efficace avec des modèles acoustiques — probabilisés par essence — qui fait leur intérêt, ou au contraire une réelle adéquation en matière de modélisation du langage parlé ? L'utilisation de modèles de langages à fort ancrage linguistique reste en tout cas une problématique ouverte en reconnaissance de parole. Il n'en va pas de même dans d'autres domaines où la pertinence des traitements linguistiques n'est plus à démontrer. Quelques exemples suffiront à étayer cette affirmation.

3.2. Etiquetage morphosyntaxique, ou le dilemme coût d'entrée / efficacité

L'étiquetage morphosyntaxique[52] consiste à attribuer à chaque mot d'un énoncé sa classe grammaticale (également appelée partie du discours ou *part of speech* en anglais). Cet étiquetage constitue fréquemment une étape préalable à l'analyse linguistique des énoncés écrits ou oraux[53]. Dans le cas de l'analyse syntaxique, il permet de réduire l'espace de recherche des parseurs en désambiguïsant les énoncés au niveau morphosyntaxique. En reconnaissance de la parole, il est possible utiliser des modèles de langages portant sur des catégories syntaxiques (grammaires N-classes). L'apprentissage de ces modèles requiert la constitution de corpus étiquetés en parties du discours.

Diverses méthodes ont été développées pour désambiguïser les parties du discours associées aux mots de l'énoncé. Celles-ci se répartissent en deux approches principales. D'un côté, on est en présence de systèmes symboliques utilisant des règles obtenues par apprentissage suivant une méthode développée par Eric Brill[54]. De l'autre, on trouve des systèmes stochastiques reposant sur l'estimation de modèles markoviens, en application directe du modèle du canal bruité[55] (cf. figure 1.3). La théorie de l'information a d'ailleurs trouvé dans cette problématique un de ses premiers champs d'application en traitement du langage naturel. Les premiers travaux sur le sujet, datés du tournant des années 1980[56], sont en effet presque contemporains des premières tentatives de reconnaissance markovienne de la parole.

La campagne d'évaluation GRACE[57] a récemment permis de comparer les performances de plus de dix étiqueteurs du français. Elle me semble suffisamment représentative pour donner une indication sur la pertinence des approches symboliques et stochastiques. Or, c'est un système à base de connaissances qui a présenté les meilleurs performances au cours de l'évaluation. Cet étiqueteur développé par Jacques Vergnes et Emmanuel Giguet au GREYC (Caen) consistait en une version, dégradée pour l'occasion, d'un analyseur syntaxique reposant sur une base de règles encodée à la

processing : a guide to theory, algorithm and system development. Prentice Hall, Upper Saddle River, NJ : p. 581-584.

52 Pour une présentation plus détaillée, on consultera : Paroubek P., Rajman M. (2000) Etiquetage morpho - syntaxique, In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection *IC2*, Hermès, Paris.

53 Valli A. et Véronis J. (1999) Etiquetage grammatical de corpus de parole : problèmes et perspectives. *Revue Française de Linguistique Appliquée*, 4(2), 113-133.

54 Les règles de la « méthode de Brill » ne doivent pas être confondues avec celles des analyseurs syntaxiques. Il s'agit en fait de règles de transformation d'étiquettes en fonction du contexte. Ces règles sont créées par apprentissage incrémental sur un corpus étiqueté : Brill E. (1992) A simple rule-based part of speech tagger. Actes 3rd *Conference on Applied Natural Language Processing*, *ANLP'1992*, Trente, Italie ; Brill E. (1995) Transformation-based error-driven learning and natural language processing : a case study in part of speech tagging, *Computational Linguistics*, 21(4), 543-565.

55 Suivant le modèle du canal bruité, l'idée est de chercher le modèle de langage qui maximise la probabilité P(S|W) d'émission d'une séquence de parties du discours S connaissant la séquence de mots W. Soit, par application de la règle de Bayes, de maximiser la probabilité P(W|S) * P(S). Les hypothèses réductrices nécessaires à la simplification du problème concernent essentiellement la limitation du contexte d'analyse (N-grammaires). Par exemple : Merialdo B. (1994) Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2), 155-172.

56 Voir les travaux sur l'étiquetage du corpus LOB : Leech G., Garside R., Atwell E. (1983) The automatic grammatical tagging of the LOB corpus. *ICAME News*, 7, 13-33 ; ou encore : Marshall I. (1983) Choice of grammatical word-class without global syntactic analysis : tagging words in the LOB corpus. *Computers and the Humanities*, 17, 139-150

57 Adda G., Mariani J., Paroubek P., Rajman M. et Lecomte J. (1999) L'action GRACE d'évaluation de l'assignation des parties du discours pour le français, *Langues*, 2(2), 119-129.

main[58]. D'autres études témoignent également du bon comportement d'analyseurs syntaxiques à base de règles sur une tâche d'étiquetage[59]. Elles suggèrent qu'une stratégie d'analyse reposant sur une désambiguïsation morphosyntaxique n'est pas obligatoirement optimale, mais également que les systèmes à base de connaissances gardent toute leur pertinence sur cette problématique. Comme le notent Paroubek et Rajman[60], les approches probabilistes conservent cependant la faveur de la majorité des chercheurs du domaine:

> *« Si l'on en juge par le nombre de systèmes disponibles actuellement, les méthodes à base de règles " pures ", c'est-à-dire sans apprentissage, semblent connaître une certaine défaveur au profit des méthodes avec apprentissage automatique, même si ces dernières semblent présenter des performances inférieures en précision d'annotation »*

On peut donner deux explications à cette situation. D'une part, les approches probabilistes bénéficient d'un effet d'entraînement dû à leurs succès dans d'autres domaines. D'autre part, il est indéniable que les méthodes à base d'apprentissage automatique — qu'il soit probabiliste ou symbolique suivant la méthode de Brill — présentent un coût d'entrée sensiblement plus faible du moment où l'on dispose des ressources linguistiques adéquates. Il est en effet plus rapide et plus facile d'obtenir un étiqueteur aux performances acceptables par apprentissage automatique que de chercher à construire manuellement un système efficace à base de règles.

Ce constat peut être étendu à d'autres domaines relevant de l'ingénierie des langues. La **traduction automatique** a également représenté une problématique pionnière pour le TAL probabiliste. De nombreuses recherches sont toujours menées dans cette perspective, comme en témoigne les efforts menés actuellement pour la constitution de larges corpus multilingues alignés[61]. Pourtant, il semble que les traducteurs probabilistes ne peuvent égaler les systèmes à base de connaissances que sur certaines applications très finalisées (traduction de bulletins météorologiques par exemple). Il n'en reste pas moins qu'il est possible d'atteindre par apprentissage automatique des performances honorables sans avoir à déployer les efforts d'équipes qui travaillent depuis des années sur des bases de règles de traduction. Philippe Langlais et ses collègues du RALI, à Montréal, n'ont ainsi jamais caché que cet élément constituait une motivation importante de leurs intéressants travaux en traduction probabiliste[62].

Cette opposition entre « coût d'entrée » prohibitif et qualité intrinsèque des méthodes à base de connaissances semble se retrouver également en **synthèse de parole à partir du texte** (*text-to-speech synthesis* en anglais). Lorsque Thierry Dutoit compare les approches à base de règles aux méthodes de synthèse par concaténation[63], il rappelle que :

> *« La conception de tels systèmes* [à base de règles] *requiert* […] *une part importante d'expérimentation par essais - erreurs, ce qui allonge le temps de développement et en alourdit le coût.* […] *Il reste que la synthèse par règles constitue en quelque sorte la voie royale pour la mise au point de voix de synthèse.* […][64] »

---

58 Cet analyseur à base d'automates à états finis s'ancre d'un point de vue théorique sur les travaux de L. Tesnière (grammaires de dépendances) : Vergnes J., Giguet E. (1998) Regards théoriques sur le tagging. Actes *TALN'1998*, Paris, France.

59 Samuelson C., Voutilainen A. (1997) Comparing a linguistic and a stochastic tagger. Actes *ACL-EACL'1997*, Madrid, Espagne.

60 Paroubek P., Rajman M. (2000) Etiquetage morpho - syntaxique, In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection *IC2*, Hermès, Paris. page 137.

61 Véronis J. (2000) Alignement de corpus multilingues. In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection I²C. Hermès, Paris, France. 152-171.

62 Langlais P. (2002) Ressources terminologiques et traduction probabiliste : premiers pas positif vers un système adaptatif. Actes *TALN'2002*, Nancy, France. 43-53 ; Langlais P., Simard M. (2003) De la traduction probabiliste aux mémoires de traduction (ou l'inverse). Actes *TALN'2003*, Batz-sur-Mer, France. 195-204.

63 La synthèse par concaténation ne repose pas sur le modèle du canal bruité, mais consiste à lisser et mettre « bout à bout » des segments de parole — déjà co-articulées — qui ont été préalablement stockés dans une grande base de données. Comme en TAL probabiliste, on retrouve une approche empirique centrée sur les données.

64 Citation extraite de (pp. 394-395) : Dutoit T. (2000) Synthèse de la parole à partir d'un texte, In Boite R., Bourlard H., Dutoit H., Hancq J. et Leich H. *Traitement de la parole*, Coll. Electricité, Presses Polytechniques et Universitaires Romandes, Lausanne, Suisse. 345-442.

Ces observations suggèrent que certaines performances des méthodes statistiques ne sauraient remettre en question l'intérêt à plus long terme des approches à base de connaissances. Certaines problématiques peuvent cependant sembler plus adaptées à des recherches statistiques. Nous allons pourtant voir que, là encore, le débat n'est pas clos.

3.3. Compréhension automatique de la parole et dialogue homme - machine finalisé

La compréhension automatique de la parole consiste à construire une représentation sémantique des énoncés obtenus en sortie d'un module de reconnaissance vocale. Elle intervient dans des applications comme le dialogue oral homme - machine, la traduction parole - parole ou encore les systèmes oraux de question - réponse[65]. Ces applications ont pour caractéristique commune une limitation de l'univers de la tâche concernée (dialogue finalisé).

La problématique de la compréhension de la parole, sur laquelle je m'attarderai dans les chapitres 3 et 4, semble adaptée à l'application de méthodes probabilistes. D'une part, elles laissent espérer un couplage optimal avec la reconnaissance de parole dans une stratégie de recherche multi-passes (cf. § 3.1). D'autre part, les modèles statistiques présentent des performances qui ne se dégradent que progressivement dans des situations atypiques. Ils semblent donc bien armés pour traiter le langage parlé et ses nombreuses structures spontanées (hésitations, répétitions, autocorrections, incises, inachèvements). Enfin, le caractère finalisé du dialogue a conduit à l'émergence de traitements sélectifs se limitant à l'identification de segments clés au sein de l'énoncé[66]. Or, il est clair que les modèles probabilistes sont pour l'heure plus adaptés à de tels traitements de surface qu'à une analyse de dépendances profondes.

Néanmoins, les résultats de diverses campagnes d'évaluation ne permettent pas de conclure à une supériorité des approches probabilistes. Les tests réalisés par la (D)ARPA sur le domaine du renseignement aérien (ATIS : *Air Transport Information Systems*) sont de ce point de vue riches d'enseignements. Le tableau 1.2. donne les performances des meilleurs systèmes ayant participé à la campagne ATIS de 1994, en regard des traitements qu'ils mettaient en œuvre[67]. On constate que les systèmes statistiques et à base de règles présentent des résultats comparables. La prédominance des modèles markoviens en reconnaissance de la parole ne saurait donc être étendue pour l'heure à la compréhension du langage parlé.

---

[65] Jamoussi S., Smaïli K., Haton J.-P. (2003) Vers la compréhension automatique de la parole : extraction de concepts par réseaux bayésiens. Actes *TALN'2003*, Batz-sur-Mer, France. 165-174.

[66] Wolfgang Minker résume ainsi le postulat sur lequel sont basées ces approches : " [*L'analyse*] *doit se limiter aux éléments porteurs de sens de la requête tout en ignorant les parties redondantes ou non-essentielles pour l'application* " : Minker W. (1999) Compréhension automatique de la parole. L'harmattan, Paris, France : 41.

[67] Pallet D.S., Fiscus J.G. et al. (1995) 1994 benchmark tests for the ARPA spoken language program. Actes *ARPA workshop on spoken language technology*. Morgan Kaufman, Princeton, NJ. 5-36.

Tableau 1.2. *— Taux de robustesse et type de modélisation mise en œuvre par les meilleurs systèmes de la campagne d'évaluation ARPA-ATIS de 1994. Résultats obtenus sur des énoncés de type A (compréhension sans contexte)*

Laboratoire (Système)	AT&T (Chronus)	CMU (Phoenix)	MIT (Galaxy)	SRI (Gemini)	BBN (Hum)
% d'erreurs	3,8	3,8	4,5	7,0	9,4
Modélisation	Statistique	Base de règles	Base de règles	Base de règles	Statistique

Cette observation se retrouve au niveau de la gestion du **dialogue homme - machine**. La proposition faite par Roberto Pieraccini et d'autres chercheurs[68] d'une modélisation stochastique de la structure du dialogue ne semble pas avoir été poursuivie. Les recherches actuelles en matière de traitement des références[69], d'identification des buts de l'utilisateur et plus généralement de contrôle du dialogue reposent ainsi généralement sur des modélisations à base de connaissances. En particulier, le dialogue a principalement donné lieu à des modélisations symboliques[70] sous forme d'approches par planification (dites aussi différentielles) ou de grammaires de dialogues (modèle hiérarchique inspiré de l'école Genevoise).

4. POUR UNE INGENIERIE DES LANGUES … PLUS LINGUISTIQUE

Ce rapide tour d'horizon nous a fourni une vision relativement équilibrée des performances respectives des approches statistiques et à bases de règles. Il est pourtant une problématique, écartée à dessein de l'état de l'art précédent, où les approches statistiques sont prédominantes. Il s'agit de l'indexation et de la recherche d'information textuelle (*text retrieval* en anglais), encore appelée recherche documentaire, voire *text mining* (fouille de texte) dans un contexte plus applicatif. Bien que ce domaine ne concerne pas mes thématiques de recherche, il me semble intéressant de s'attarder sur ses réussites et surtout sur les problèmes qu'il rencontre. Ceux-ci me semblent en effet révélateurs des difficultés auxquelles peut conduire une pratique ingénierique pas assez soucieuse du fait linguistique. Problème que nous allons chercher à cerner et auquel tente de répondre le programme scientifique que je conduis avec mon groupe de recherche.

4.1. Un cas d'école : la recherche d'information textuelle

La recherche d'information textuelle consiste à fournir un ensemble de documents pertinents, voire une information précise[71] extraits d'une grande banque de données textuelles en réponse à une requête. Lorsque cette question et la réponse peuvent être exprimée sous la forme d'énoncés en langage naturel, on parle de système de question / réponse (*question / answering* en anglais). La

[68] Roberto Pieraccini et Esther Levin proposent une détermination stochastique de la stratégie de dialogue à adopter en fonction du contexte. On citera également les travaux de Reithinger et Klesen sur la caractérisation probabiliste des actes de dialogue : Pieraccini R., Levin E. et Eckert W. (1998) Spoken Language Dialogue: architecture and algorithms. Actes *XXII° Journées d'études sur la parole*, *JEP'98*, Martigny, Suisse, 387-395 ; Levin E. et Pieraccini R. (1997) A stochastic model of computer-human interaction for learning dialogue strategies. Actes *5th European Conference on Speech Communication and Technology*, *Eurospeech'97*. Rhodes, Grèce. 1883-1886 ; Reithinger N. et Klesen M. (1997), Dialogue act classification using language models. Actes *5th European Conference on Speech Communication and Technology*, *Eurospeech'97*. Rhodes, Grèce. 2235-2238.

[69] Pierrel J-M., Romary L. (2000) Dialogue homme - machine. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 331-350 ; Mitkov R., Boguraev B., Lappin S. (Eds.) (2001) Special issue on computer anaphora resolution. *Computational Linguistics*, 27(4).

[70] Moeschler J. (1989) Modélisation du dialogue. Hermès, Paris, France ; Litman D. J. (1985) Plan recognition and discourse analysis : an integrated approach for understanding dialogues. Thèse de Doctorat, U. de Rochester, Etats-Unis ; Guyomard M., Nerzic P., Siroux J. (1993) Plan, méta plans et dialogue. Actes de la *4ème école d'été sur les traitements des langues naturelles*. Lannion, France ; Pierrel J-M. (1987), Dialogue oral homme - machine. Hermès, Paris, France ; Alexandersson J., Reithinger N., Maier E. (1997), Insights into the dialogue processing of VERBMOBIL, Actes *5th Conference on Applied NLP*, *ANLP'97*, Washington, DC, 33-40.

[71] Les systèmes participant à la campagne TREC-10 (*Text Retrieval Conference*) sont évalués uniquement sur une réponse fournie par question posée. Jusqu'ici, les campagnes d'évaluation TREC-8 et TREC-9 reposaient sur la sélection de courts extraits de documents (50 ou 250 caractères suivant les cas).

recherche consiste à estimer la pertinence des documents de la banque de données en les comparant à la question. Ce problème se résume donc à un calcul de similarité entre éléments. Le nombre de documents à explorer et à classer est très important. Leur filtrage est donc d'autant plus fin qu'il est envisagé dans un espace numérique continu et non dans un cadre booléen[72]. Le problème de la recherche d'information peut donc être reformulé comme suit :

a) définir un système de représentation qui permet de décrire tout document ou question dans un espace quantifiable numériquement,

b) choisir ensuite une métrique qui permet de calculer une distance entre objets dans l'espace défini par le système de représentation.

Ainsi reformulée la recherche d'information apparaît comme une problématique dédiée à un traitement numérique plus qu'à une approche symbolique. De fait, les chercheurs se sont intéressés dès le début des années 1970 à des approches statistiques permettant une comparaison pondérée question / documents. Compte tenu de la nature du problème, ces approches relevaient de techniques classificatoires et non pas du modèle du canal bruité. Elles revêtent donc un caractère particulier en ingénierie des langues.

Parmi ces approches numériques, c'est le modèle vectoriel élaboré par Gérard Salton en 1972 qui a permis une avancée significative des recherches[73]. Celui-ci a proposé comme système de représentation un espace vectoriel construit sur un ensemble de mots clés jugés significatifs. Chaque question ou document est donc représenté par un vecteur. Ses composantes sont estimées suivant une statistique plus ou moins évoluée. Celle-ci repose généralement sur la fréquence d'apparition dans le document du mot clé tenant lieu de vecteur directeur, relativement à la fréquence du mot dans l'ensemble de la banque de données[74]. Cette représentation vectorielle étant définie, plusieurs métriques peuvent être utilisées pour estimer la similarité entre la question et les documents à filtrer. La plus utilisée correspond au cosinus de l'angle entre les deux vecteurs respectifs.

Comme on le voit, le modèle vectoriel et ses successeurs tels que l'approche LSI (*Latent Semantic Indexing*) sont purement statistiques. La seule connaissance linguistique résulte du choix des mots clés qui définissent les vecteurs directeurs de l'espace vectoriel. Cette approche adaptée à des calculs intensifs sur de grandes masses de données a rapidement dominé les travaux du domaine. Dans un article dressant un état de l'art de la recherche documentaire en 1995, Donna Harman et ses collègues précisent ainsi l'état d'esprit général de la communauté scientifique sur l'opposition statistique / symbolique[75] :

> *« Most researchers in the information retrieval community believe that retrieval effectiveness is easier to improve by means of statistical methods than by NLP-based approaches and this is borne out by results, although there are exceptions. The fact that only a fraction of information retrieval research is based on extensive natural langage processing techniques indicates that NLP techniques do not dominate the current thrust of information retrieval research as does something like the Vector Space Model »*

---

[72] S'ils utilisent des scores pondérés pour ordonner leurs résultats, les outils de recherche sur la Toile en restent encore à la satisfaction d'une requête booléenne portant sur plusieurs mots - clés.

[73] Salton G. (1972) Experiments in automatic thesaurus construction for information retrieval. Actes congrès de l'*IFIP'1972*, Ljubljana, Slovénie. Communication citée dans : Fluhr C. (2000) Indexation et recherche d'information textuelle. In Pierrel J-M. (Dir.) Ingénierie des langues. *Collection I²C*. Hermès, Paris, France. 235-252.
Pour un exposé complet sur modèle vectoriel et ses évolutions : Salton G. (1989) Automatic text processing. The transformation, analysis and retrieval of information by computer. Addison-Wesley, New-York, NJ.

[74] Dans l'évolution la plus développée du modèle vectoriel — l'approche LSI (*Latent Semantic Indexing*) — cette statistique n'est plus aussi directe. On cherche en effet à réduire a posteriori l'espace vectoriel (décomposition en vecteurs propres de la matrice termes de base / documents de la banque de données) pour rendre compte de relations sémantiques implicites entre les termes tenant lieu de vecteurs de base. Le principe fondateur de la méthode reste cependant le même : Deerwester S., Dumais S., Landauer T. Furnas G., Harshman R. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*. 416(6), 391-407.

[75] Harman D., Schäube P., Smeaton A. (1995) Document retrieval. *In* Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon. Disponible sur la Toile : http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html (citation p. 261).

Pourtant, il ne fait guère de doute que la recherche d'information textuelle met en jeu chez l'être humain des fonctions cognitives de lecture et de compréhension qui reposent sur l'utilisation de connaissances linguistiques variées. Sans suivre une approche cognitiviste, quelques exemples suffisent à démontrer l'intérêt de l'intégration d'informations linguistiques dans le processus de recherche d'information[76] :

- **résolution de l'ambiguïté lexicale et analyse syntaxique de la requête** — Afin de bien identifier le sens de la requête de l'utilisateur, mais aussi celui des documents de la base, il est nécessaire de lever toute l'ambiguïté lexicale sur les éléments qui les composent. Il faut donc pouvoir caractériser le lemme correspondant à chaque mot. Dans des langues à morphologie relativement riche, ce problème ne peut être résolu sans le recours à une véritable analyse linguistique. Souvent, la détermination de la catégorie syntaxique du lemme par un étiqueteur grammatical est nécessaire. Elle permet de résoudre de multiples cas d'homographies tels que le célèbre *couvent* (verbe couver *conjugué* ou nom commun synonyme de *monastère*) en français, sans parler des problèmes de désaccentuations[77] ou d'identification des mots composés tels que les termes en *N de N*, beaucoup moins ambigus que leurs composants monolexicaux[78].

  Les méthodes statistiques à faible ancrage linguistique devront disposer d'une masse considérable de données pour pouvoir espérer distinguer certaines des ambiguïtés citées. En tout état de cause, cette taille significative ne peut être atteinte par la question posée par l'utilisateur. Cela explique que les recherches actuelles mettent principalement l'accent sur l'analyse de la question posée. L'apport de la linguistique ne s'y limite pas à la désambiguïsation lexicale de la question posée. Bien souvent, le recours à une analyse syntaxique complète peut éviter de lancer la recherche dans une voie erronée. Dans un article récent[79], Laura Monceaux et Isabelle Robba nous donnent deux exemples éclairants sur des tests issus de la campagne TREC-9 (*Text Retrieval Conference)* : en l'absence de segmentation en constituants noyaux (*chunks*) dans un cas, ou d'analyse syntaxique précise permettant de déterminer les relations prédicats - arguments dans un second, la recherche est vouée dès le départ à une erreur pourtant triviale.

- **expansion de requête : apport de la sémantique lexicale** — Rechercher un document, c'est d'abord le comprendre. Aussi ne s'étonnera-t-on pas que l'utilisation d'informations sémantiques puisse améliorer le processus de recherche. A l'heure actuelle, cet apport est envisagé une fois encore au niveau de la requête. L'idée est d'utiliser certaines relations sémantiques entre mots pour étendre (ou reformuler) la question posée, suivant un processus relativement intuitif : si vous cherchez des informations sur les *voitures*, il est clair que vous serez intéressés par tout document portant sur des *automobiles* (relations de synonymie), voire sur les *véhicules* en général (relations d'hyponymie / hyperonymie). On assiste ainsi au développement de grandes ressources lexicales telles que Wordnet[80] ou Eurowordnet dont l'objectif est de constituer de vastes ontologies couvrant les principales relations utilisées en sémantique lexicale. Pour l'heure, l'utilisation de ces ressources ne semble pas avoir donné les résultats escomptés. Cette situation semble cependant devoir être attribuée à l'inadéquation des ressources utilisées (ontologie trop générale) pour la tâche considérée. Pierrette Bouillon et ses collègues montrent ainsi l'intérêt d'une ressource lexicale apprise sur un corpus du domaine, donc spécifique à ce dernier[81].

---

76 Pour une présentation moins succincte de l'apport de la linguistique en recherche d'information textuelle, on lira le texte de synthèse de Christian Fluhr : Fluhr C. (2000) op. cit.

77 Zweigenbaum P., Garbar N. (2002) *Accentuation de mots inconnus : application au thesaurus biomédical MeSH*. Actes TALN'2002, Nancy, France. 53-62.

78 Dias G., Guillore S., BassanoJ.-C., Pereira Lopes J. C. (2000) Extractions automatiques d'unités lexicales complexes : un enjeu fondamental pour la recherche documentaire. *TAL*, vol. 41 n° 2. 447-472. Hermès, Paris, France.

79 Monceaux L., Robba I. (2002) Les analyseurs syntaxiques : atouts pour une analyse des questions. Actes *TALN'2002*, Nancy, France. 195-204.

80 Fellbaum C. (1998). Wordnet, an electronic database. MIT Press, Cambridge, MA.

81 Bouillon P., Fabre C., Sébillot P., Jacqmin L. (2000) Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL*, vol. 41 n° 2. 447-472. Hermès, Paris, France. 367-393.

L'apport des connaissances linguistiques à la recherche d'information textuelle semble avoir été relevée dès les années 1990. L'article de synthèse de Donna Harman et ses collègues rappelle ainsi le foisonnement des tentatives faites dans cette direction, en particulier dans le cadre des projets TIPSTER[82]. Bien souvent, ces propositions visaient le cœur de la recherche d'information, à savoir la méthode d'appariement question / document. Les résultats obtenus furent cependant décevants. On constate en particulier la difficulté qu'il y a à combiner approches statistiques et traitements linguistiques[83] :

> *« There is* […] *an inherent granularity mismatch between the statistical techniques used in information retrieval and the linguistic techniques used in natural language processing*. […] *Further refinement of this process* [of statistical matching] *using fine-grained natural language processing techniques often had only noise* […] »

Pour se sortir de cette impasse, les chercheurs ont alors suivi une stratégie que je qualifierais de démarche centrée ingénierie, à laquelle répondait les campagnes de test TREC naissantes. Cette stratégie consiste à ne toucher qu'à la marge le processus central d'appariement, qui avait permis des avancées significatives, tout en effectuant au préalable des traitements linguistiques souvent fouillés sur la question et les documents.

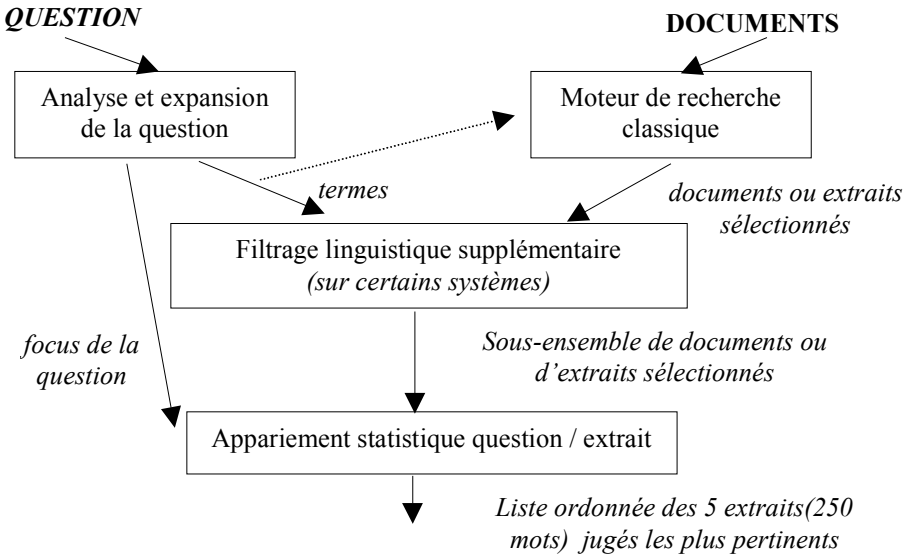

Figure 1.5 *— Architecture générique d'un système de recherche d'information textuelle de type question / réponse[84].*

A titre d'illustration, la figure 1.5. présente l'architecture générique de la plupart des systèmes question / réponse qui ont participé à la campagne TREC-9 (2000). Celle-ci repose sur un traitement séquentiel de l'information. Dans un premier temps, on procède à un traitement linguistique assez fin de la question posée. Cette étape a deux finalités. D'une part, on cherche à déterminer le focus de la question afin de guider ultérieurement la recherche de la réponse. D'autre part, une analyse à composante syntaxique[85] vise à déterminer sans ambiguïté les éléments saillants de la question. Ceux-ci sont ensuite étendus (recherche de synonymes, d'hyperonymes) à l'aide d'une connaissance sémantique lexicale. Les termes extraits de la requête étendue permettent de lancer un moteur de

82 Harman D., Schäube P., Smeaton A. (1995) op. cit.

83 Harman D., Schäube P., Smeaton A. (1995) op. cit. (citation p. 262).

84 Cette architecture est inspirée du système QALC du LIMSI-CNRS (Orsay, France). Comme l'indiquent les auteurs, celle-ci se retrouve dans les autres systèmes participant à TREC-9 : Ferret O., Grau B., Hurault-Plantet M., Illouz G., Jacquemin C. (2001) Utilisation des entités nommées et des variantes terminologiques dans un système question - réponse. Actes *TALN'2001*, Tours, France. 153-162.

85 On procède ici généralement à un étiquetage morphosyntaxique de la question, voire à une analyse syntaxique de surface (*shallow parsing*). Voir l'article de Laura Monceaux et Isabelle Robba (*op. cit*.) sur ce dernier point.

recherche traditionnel sur la base de documents. On récupère alors plusieurs centaines de documents supposés pertinents, qui peuvent être filtrés de nouveau par des traitements linguistiques supplémentaires. Selon Olivier Ferret et ses collègues[86], cette étape conduit à une excellente précision, de l'ordre de 95%.

Enfin intervient l'étape d'appariement entre la question et les documents (ou leur extraits sous forme de phrases dans le cas de TREC). Les principes du modèle vectoriel sont utilisés ici avec très peu de modifications. On notera simplement la prise en compte du focus de la question sous forme de pondérations des termes correspondant aux entités recherchées. De même, les termes utilisés pour définir l'espace vectoriel de recherche sont pondérés suivant leur présence effective dans la question ou dans l'expansion de cette dernière.

La recherche finale de la réponse repose toujours sur un appariement statistique sans ancrage linguistique. Il me semble alors difficile de conclure comme Christian Fluhr à une intégration réussie entre traitements linguistiques et approches ingénieriques propres à traiter de grandes masses de données[87]. Sur ce point, Pierrette Bouillon et ses collègues sont claires[88] :

> *« Dans leur grande majorité, les systèmes de recherche d'information n'exploitent pas d'informations de nature linguistique* »

Cette situation a certainement pour origine la difficulté que représente la mise en place de méthodes linguistiques d'appariement fiables. On peut également y voir l'influence des campagnes de tests TREC qui, en se répétant chaque année, favorisent la maîtrise des processus d'ingénierie au détriment de l'innovation scientifique. Nous reviendrons d'ailleurs (chap. 3 § 1) sur les apports et travers des paradigmes d'évaluation. Quoi qu'il en soit, la recherche documentaire semble avoir privilégié le recours à des solutions ingénieriques qui sont certes éprouvées mais qui offrent peu de perspectives par manque d'ancrage linguistique.

---

86 *op. cit*. Citation page 157. Dans le cas du système QALC du LIMSI, c'est le moteur Indexal de la société Bertin Technologies qui a été utilisé : de Loupy C., Bellot P., El-Bèze M., Marteau P.-F. (1998) Query expansion and classification of retrieval documents. Actes *Text Retrieval Conference, TREC-7,* Gaithersburg, Maryland. 382-389.

87 Fluhr C. (2000) op. cit. (p. 250).

88 Bouillon P. *et al*. (2000) op. cit. (p. 368).

Tableau 1.3. *— Campagne de test TREC-9 (2000). Résultats bruts et pondérés (tenant compte du rang de la réponse attendue parmi 5 propositions) des meilleurs systèmes. Résultats sur une réponse en 250 caractères*[89]

Laboratoire (système) (Auteurs)	(FALCON) Harabagiu *et al*.[90]	(PIRCS) Kwok *et al*.[91]	IBM Ittycheriah *et al*[92].	LIMSI (QALC) Jacquemin *et al*.[93]
Rang TREC'9 (250 caractères)	1er	2ème	4ème	6ème
% réponses correctes (parmi 5)	77,8%	67,3%	56,2 %	55,00%
Score pondéré	0,760	0,460	0,457	0,407

Les résultats de TREC'9 sont de ce point de vue révélateurs des limites de cette stratégie « centrée ingénierie ». Le tableau 1.3 donne un aperçu des résultats présentés à la conférence TREC-9 (2000). L'évaluation consistait à fournir, pour 200 questions posées, 5 réponses ordonnées extraites d'un corpus d'un million de documents issus de journaux anglophones. Les réponses devaient être formulées sous la forme de courts extraits (250 caractères maximum) des documents. Cette évaluation repose sur un scénario proche de l'activité réelle de recherche documentaire. Les requêtes de TREC-9 proviennent d'ailleurs de questions réelles. Du point de vue de l'utilisateur, ces résultats me semblent donc plus parlants que des métriques usuelles telles que les taux de rappel et de précision[94].

Ces résultats se caractérisent par une domination assez marquée du système classé premier sur ses suivants immédiats. Les systèmes classés entre la seconde et la sixième position sont parvenus à fournir la bonne réponse — parmi quatre autres réponses éventuellement erronées — dans environ 60 % des cas. Nos multiples furetages sur la Toile nous ont accoutumé à une précision plus faible. Il n'en reste pas moins que ces résultats restent largement perfectibles. En particulier, le système classé premier (FALCON) a réussi cette tâche dans près de 78% des situations[95]. Cette domination est encore plus marquante si l'on s'intéresse aux scores pondérés (0,76 contre 0,46 pour le second) qui prennent en considération le rang de la bonne réponse parmi les propositions des systèmes.

Or, la principale différence entre ces systèmes réside dans l'étape finale d'appariement entre la question et les documents. Là où les autres systèmes restent proches du modèle vectoriel, FALCON[96] renouvelle la problématique en utilisant une approche sémantique. Il réalise une unification entre la représentation sémantique de la question et celle des extraits de documents sélectionnés.

---

89 Pour plus de renseignements sur les résultats de la campagne de test, on peut consulter les actes de la conférence disponibles sur la Toile : http://trec.nist.gov/pubs/trec9/t9_proceedings.html

90 Harabagiu S, Moldovan D. (2000) FALCON : boosting knowledge for answer engines. Actes *Text Retrieval Conference*, *TREC-9*. Gaithersburg, Maryland, USA. 479-488.

91 Kwok K.L., Grunfeld L., Dinstl N., Chan M. (2000) TREC-9 Cross Language, Web and Question-Answering Track Experiments using PIRCS. Actes *Text Retrieval Conference*, *TREC-9*. Gaithersburg, Maryland, USA. 419-428.

92 Ittycheriah A., Franz M., Zhu W.J., Ratnaparkhi A. (2000) IBM Statistical Answering System. Actes *Text Retrieval Conference*, *TREC-9*. Gaithersburg, Maryland, USA. 229-238.

93 Ferret O., Grau B., Hurault-Planter M., Illouz G., Jacquemin C. (2000) QALC : the question-answering system of LIMSI-CNRS. Actes *Text REtrieval Conference*, *TREC-9*, Gaithersburg, Maryland, USA. 235-244.

94 Ces mesures sont néanmoins très utilisées en recherche documentaire. Celles-ci sont définies comme suit. On note C le nombre de réponses pertinentes fournies par le système, R le nombre de documents totaux (pertinents ou non) proposés par le système, et enfin D le nombre de documents pertinents qui auraient du être proposés. On définit alors les mesures suivantes :
rappel = C / D et silence = 1 – rappel
précision = C / R et bruit = 1 – précision

95 Plus précisément, le taux de réussite du système FALCON a été de 77,8 % sur une réponse longue (250 caractères) et de 59,9 % sur réponse courte (50 caractères).

96 Harabagiu S., Pasca M., Maiorano J. (2000) Experiments with open-domain textual question answering, actes *COLING'2000*, Saarbrücken, Allemagne. 292-298.

Cette remise en cause du paradigme dominant de la recherche d'information explique-elle la réussite du système FALCON ? Cela semble être l'opinion des observateurs du domaine. Comme le remarquent Olivier Ferret et ses collègues[97] :

> *«* [La] *réponse étant recherchée dans une grande masse de données, il est tentant d'appliquer des méthodes essentiellement numériques pour la trouver. Néanmoins, les expériences montrent que l'ajout de raisonnements fondés sur des connaissances sémantiques et pragmatique est nécessaire* […] *En effet, le système qui a obtenu les meilleurs résultats est celui qui utilise le plus largement les techniques d'analyse syntaxique et sémantique* »

Il y a peut-être là matière à une (r)évolution paradigmatique du domaine. En tous cas, l'exemple de la recherche documentaire fournit quelques indications sur les limites d'une démarche purement ingénierique qui méconnaîtrait la nature linguistique des objets sur lesquels elle travaille.

4.2. Dialogue oral homme - machine : pour une ingénierie plus linguistique

Ce constat est susceptible de concerner l'ensemble du traitement automatique des langues, dont les recherches reposent sur une démarche ingénierique de plus en plus affirmée. Il me semble utile de relever dans ce chapitre introductif quelques problèmes généraux auxquels est confrontée l'ingénierie des langues. Ces limitations seront présentées ici sans entrer dans les détails de l'argumentation. Elles seront discutées dans les chapitres qui suivent à la lumière de mes travaux en dialogue homme-machine.

4.2.1. Critères prédictifs pour la conduite des recherches en ingénierie des langues

Toute démarche ingénierique consiste à trouver une solution optimale à un problème donné, sans ce soucier nécessairement de l'intelligibilité de cette solution. L'intérêt de cette approche est de privilégier la confrontation au problème à des considérations théoriques qui peuvent parfois reposer sur des présupposés subjectifs. On conçoit cependant rapidement les limites de cette démarche.

Par exemple, manipuler des données sans chercher à les comprendre ne fournit que peu d'indices prédictifs pour déterminer la pertinence à long terme d'une solution. Il est certes possible de suivre une méthodologie de conception qui permet d'améliorer les performances des systèmes par des tests de validation réguliers. Ces progrès risquent cependant d'être marginaux. En guise d'analogie, on pourrait dire qu'une démarche purement ingénierie nous garantit d'arriver à un optimum local de manière efficace, mais que rien ne nous assure de la qualité de cette solution par rapport à un fonctionnement attendu. Le caractère faiblement prédictif de cette démarche risque en outre de nous fournir bien peu d'éléments pour imaginer une solution alternative[98].

Dans le cas de l'ingénierie des langues, cette logique a rarement été poussée à l'extrême. Certes, l'entraînement sur corpus des modèles de langage repose sur un processus d'apprentissage relativement aveugle. Dans son activité quotidienne, le chercheur recourt néanmoins à des analyses de sessions d'utilisation (*logfiles*) pour comprendre les erreurs de son système. La nature linguistique du matériau traité y est considérée implicitement. Cette validation, spécifique au système développé, ne permet cependant pas de juger de la pertinence d'une solution par rapport à une autre. L'ingénierie des langues s'est certes développée autour de plusieurs campagnes d'évaluation comparative d'envergure. Malheureusement, les métriques de test retenues se limitent le plus souvent à une estimation globale du degré de réalisation de la tâche (taux d'erreurs généraux). Je chercherai à montrer au cours du chapitre 3 que ces évaluations, qui ne fournissent aucun diagnostic linguistique, sont peu utiles à l'amélioration des systèmes.

---

97 *op. cit.* Citation page 161.

98 On peut se demander si les techniques connexionnistes ne font pas face actuellement à ce dilemme. Après avoir montré leur efficacité dans certains domaines, les perceptrons multi-couches peinent en effet à s'attaquer à d'autres problématiques. Or, il est vrai qu'un réseau de neurones constitue le plus souvent une boîte noire dont les états internes sont très peu explicables. Il en résulte un manque pénalisant d'intelligibilité auquel ne répond pas les diverses propositions d'architectures complémentaires (réseaux récurrents par exemple) élaborées ces dernières années.

4.2.2. Prise en compte de l'utilisateur final

En se fixant comme objectif principal la réalisation de systèmes opérationnels, l'ingénierie des langues semble bien armée pour intégrer l'utilisateur final dans sa réflexion. En particulier, la conception des systèmes interactifs répond généralement à la méthodologie du cycle de vie en spirale définie par le génie logiciel. Cette méthodologie de développement et test de prototypes successifs met l'utilisateur — ou du moins la tâche à satisfaire — au centre des préoccupations. Son application à l'ingénierie des langues ne va cependant pas sans poser quelques questions.

Tout d'abord, on fera remarquer que l'on ne connaît encore que bien imparfaitement les attentes de la société en matière de technologies langagières. La place encore marginale qu'elles occupent dans notre vie quotidienne explique certainement largement cette méconnaissance. Il est d'ailleurs à prévoir que la généralisation de certaines applications se traduira par des détournements d'usages, comme ce fut le cas par le passé pour bien d'autres nouvelles technologies. Comme le rappellent Jean Caelen et ses collègues au sujet des systèmes de dialogue oral[99] :

> *« On se situe dans un contexte non stabilisé d'utilisation car les systèmes de dialogue homme – machine sont à la fois innovants et à la fois concurrents par rapport à d'autres systèmes interactifs* […] *Ces systèmes n'ont pas encore une signification d'usage (on ne sait pas bien non plus à quels types d'usages ils s'adressent). »*

Par ailleurs, il n'est pas inutile de rappeler que les expérimentations menées en ingénierie des langues ne concernent le plus souvent pas l'utilisateur lambda. Le recrutement de sujets étant assez difficile, on s'en remet souvent à des familiers du domaine (étudiants, proches ou connaissances), voire dans le cas le plus favorable, à des personnes fortement motivées. Cette motivation peut provenir d'une prédisposition personnelle pour les nouvelles technologies. C'est le cas des expériences où le recrutement fait appel à une démarche volontaire des sujets (communication d'un numéro de téléphone ou d'un site Internet à contacter), ou lorsque ceux-ci ont la liberté d'accepter ou refuser l'expérimentation (choix proposé lors de la prise de contact avec un centre d'appel). Plusieurs études en sociologie cognitive ont montré que cette prédisposition, de même que le sentiment de libre arbitre, influent de manière significative sur le comportement des utilisateurs[100].

Dans d'autres cas, cette motivation est obtenue par une rétribution financière symbolique servant à valoriser l'enjeu de l'expérience. Comme l'ont noté les expérimentateurs, cette rémunération influe sur l'implication des sujets, et donc sur la qualité des données recueillies[101]. Malheureusement, nous sommes alors dans une situation peu écologique d'utilisation.

Au final, ces expérimentations sont menées sur des situations peu représentatives de celles auxquelles est a priori destiné le système[102]. Ces faiblesses ne sauraient être imputées à une démarche scientifique particulière. Il me semble cependant qu'une approche purement ingénierique ne peut qu'aggraver cette méconnaissance de l'usager final.

Rappelons tout d'abord qu'il est souvent difficile de caractériser le comportement linguistique des utilisateurs. Comme l'a montré par exemple Françoise Gadet, tout locuteur maîtrise plusieurs registres de langues auxquels il peut recourir simultanément dans une même situation[103]. Cette

---

[99] Caelen J., Zeiliger J., Bessac M., Siroux J., Perennou G. (1997) Les corpus pour l'évaluation du dialogue homme - machine. Actes des *1ères Journées Scientifiques et Techniques FRANCIL*, *JST'1997*, Avignon, France, 215-222. Texte repris dans : Chibout K., Mariani J., Masson N., Néel F. (Dir.) (2000) Ressources et évaluations en ingénierie des langues. De Boeck Université, Duculot, Bruxelles, Belgique. 417-435 (citation page 418).

[100] Fisher-Lokou J., Guéguen N., 2001, Impact of a mediator, mutual representation of the negociators and decision making in a dyad : evaluation in the case of computer-mediated-communication, *Studia Psychologica*, 43(1), 13-21.

[101] Rosset S. (2000) Stratégies et gestionnaire de dialogue pour des systèmes d'interrogation de bases de données à reconnaissance vocale. Doctorat Université Paris XI, Orsay, France. publié comme rapport de recherche 2001-18 du LIMSI-CNRS, Orsay, France. septembre 2001 (p. 210-211)

[102] On ne parlera pas en outre du problème social que représentent les personnes exclues des nouvelles technologies (personnes âgées, personnes à faible niveau de formation ou issues de quartiers défavorisés).

[103] Gadet F. (1999) La variation diaphasique en syntaxe. In Barbiéris J.-M. (Ed.) Le français parlé : variété et discours. *Praxiling*, Université de Montpellier III. 211-228.

variation diaphasique concerne le traitement automatique des langues dans sa généralité et se retrouve en particulier en dialogue oral homme - machine[104].

Cette capacité d'adaptation est généralement interprétée comme un facteur propice au développement des technologies langagières. Elle suggère en effet que l'utilisateur peut s'adapter aux contraintes qui lui sont imposées. Dans le cas des méthodologies de conception incrémentale utilisées par l'ingénierie des langues, elle me semble être au contraire un frein à la satisfaction des besoins réels de l'utilisateur. Il est en effet à craindre que ce dernier n'adapte son comportement pour contourner les insuffisances du système dont il a pris conscience. Cette adaptation palliative, bien connue en dialogue homme - machine, peut masquer des problèmes limitant fortement l'utilisabilité des systèmes. Elle n'est pas sans coût cognitif ou social. Dans une situation naturelle (non imposée) d'interaction, l'utilisateur pourra estimer ce coût plus important que l'avantage que lui procure la technologie proposée. Ce refus, qui représente un enjeu économique évident, n'a rien d'un cas d'école. On le retrouve par exemple chez ces nombreux usagers des transports en commun qui préfèrent patienter au guichet plutôt que s'en remettre à un distributeur automatique dont ils semblent pourtant maîtriser l'utilisation.

Il n'est certes pas question de rejeter toute cycle de validation au motif des capacités d'adaptation de l'être humain. Simplement, il me semble que la prise en compte de l'utilisateur ne saurait être limitée à une vérification *a posteriori* de l'utilisabilité des prototypes. Il faut au contraire chercher à mieux connaître en amont les besoins des utilisateurs. Pour cela, on analysera de manière détaillée leurs usages langagiers. Ceux-ci seront étudiés sur des situations réelles ne mettant pas forcément en jeu le système ou sa simulation. Sinon, comment parler d'analyse des besoins lorsque celle-ci se limite, en forçant un peu le trait, à la collecte de données (corpus d'entraînement) auquel le système devra se conformer aveuglément ? Je chercherai à montrer dans le chapitre suivant l'intérêt pour l'ingénierie des langues de ces analyses d'usages. Celles-ci fondent la démarche scientifique du groupe de recherche que je dirige.

De même, comment juger de la satisfaction des besoins lorsque la validation se limite au calcul d'un taux d'erreur global faisant fi de toute analyse préalable des usages langagiers ou de l'opinion des sujets ayant réalisé l'expérimentation ? Certes, plusieurs paradigmes d'évaluation actuels reposent sur des critères subjectifs tels que la satisfaction de l'utilisateur. Des propositions ont été faites en direction de leur objectivation partielle. Aussi intéressantes soient-elles, elles restent cependant trop générales pour fournir un diagnostic précis des insuffisances qualitatives des systèmes. Couplée à une analyse préalable des usages langagiers, une évaluation qui viserait un diagnostic fin du comportement du système devrait permettre de mieux répercuter à tous les niveaux de traitement les besoins des utilisateurs finaux. Nous reviendrons sur cette question au cours du chapitre 3 consacré à nos travaux sur l'évaluation des systèmes de compréhension.

4.2.3. Manque de généricité des recherches en ingénierie des langues

La notion de généricité couvre des problèmes multiples. Dans le cadre du dialogue homme - machine finalisé, elle désigne en premier lieu la portabilité du système d'une tâche ou d'un domaine d'application à un autre. Elle nous interroge également d'une manière plus générale sur l'indépendance des systèmes vis-à-vis du genre langagier (productions libres ou contraintes, langage général ou de spécialité, etc.), de l'idiome ou encore du type d'utilisateurs visés (professionnels ou grand public, novices ou entraînés, personnes âgées, handicapés, etc.).

Ce dernier point recoupe la prise en compte de l'utilisateur évoquée précédemment. Dans ce cas, la généricité peut être envisagée de deux manières. Soit le système est capable de gérer a priori tout type d'utilisateurs. C'est le cas par exemple des systèmes de reconnaissance de parole multi-locuteurs. Soit le système peut s'adapter à chaque usager. Cette capacité, qui se retrouve dans les systèmes de dictée vocale à adaptation dynamique, est essentielle pour les applications dédiées au monde du handicap (cf. chap. 4 § 2).

[104] Gufstafson J., Larsson A., Carlson R., Hellman K. (1997) How do system questions influence lexical choices in user answers ? Actes *5th European Conference on Speech Communication and Technology*. *Eurospeech'97*, Rhodes, Grèce. 2275-2278.

Enfin, il faut distinguer la généricité intrinsèque montrée du système de celle des méthodes sur lesquelles il repose. Par exemple, un système de compréhension de parole développé pour une langue donnée peut être porté dans un autre idiome[105], sans que l'on ait à modifier les techniques sous-jacentes.

D'un point de vue ingénierique, la portabilité et la réutilisabilité des systèmes constituent des facteurs de qualité essentiels. Leur importance économique n'est d'ailleurs pas à démontrer. La généricité ne saurait être cependant limitée à ces aspects purement technologiques. Au contraire, elle interroge l'ingénierie des langues dans ses fondements et ses pratiques actuelles.

Tout d'abord, les campagnes d'évaluation qui servent à juger la pertinence d'approches alternatives nécessitent des efforts conséquents de la part des concepteurs de systèmes. C'est pourquoi on ne peut les multiplier sur différents contextes. On ne dispose donc le plus souvent que d'indications spécifiques qu'il semble aventureux d'extrapoler en résultats plus généraux. En particulier, les campagnes d'évaluation qui ont été menées jusqu'à présent en dialogue homme-machine n'ont concerné qu'un nombre limité de domaines d'applications assez particuliers[106].

Cette situation peut être imputée à un manque de moyens et non à une limitation intrinsèque de l'ingénierie des langues. On relèvera néanmoins que les campagnes de test ingénieriques favorisent la recherche d'une adéquation optimale à la tâche choisie, au détriment de considération à plus long terme. D'où un risque de confusion entre la pertinence générale d'une solution et une spécialisation réussie sur une tâche donnée. Certains résultats obtenus dans le cadre des campagnes TREC sont de ce point de vue éloquents. A titre d'exemple, Christian Jacquemin et ses collègues du LIMSI rappellent que leur système a été classé respectivement 6ème et 28ème lors de la campagne TREC-9 (2000), suivant la longueur de la réponse autorisée (50 et 250 caractères) [107]. Comment appréhender la pertinence générale d'une solution sur des performances si différentes pour des tâches pourtant proches ?

Par ailleurs, il est envisageable de proposer une alternative à la réalisation de multiples campagnes d'évaluation sur des domaines différents. Il s'agit de définir en amont les usages langagiers (caractérisation syntaxique du langage, études pragmatiques des structures de dialogues, etc.) de chaque contexte applicatif et de disposer d'un diagnostic détaillé du comportement des systèmes à mettre en regard de cette caractérisation. Moyennant certaines extrapolations, on peut alors espérer prédire, même imparfaitement, l'adéquation d'une technique donnée à une nouvelle classe de problème. On disposerait alors d'un diagnostic prédictif et générique sur la solution envisagée.

4.2.4. Conclusion : un programme de recherche centré sur le fait linguistique

Ce survol de certaines difficultés rencontrées par le traitement automatique des langues ne saurait masquer l'apport des approches ingénieriques. L'analyse de ces insuffisances m'a simplement conduit à adopter et défendre le principe d'une démarche technologique plus ancrée sur des considérations linguistiques.

Je vais précisément m'intéresser maintenant à mes travaux de recherche, en montrant sur des cas précis en quoi une démarche plus soucieuse de la réalité linguistique peut répondre à ces interrogations. Dans un premier temps, je m'intéresserai aux conséquences méthodologiques de cette démarche, en présentant successivement mes travaux en linguistique de corpus (chapitre 2) et

[105] A titre illustratif, on citera les portages anglais / français réalisés par le laboratoire LIMSI : Minker W. (1995) An English version of the LIMSI L'ATIS System. Rapport technique LIMSI 95-12, Orsay, France ; Bonneau-Maynard H., Gauvain J.-L., Goodine D., Lamel L., Polifroni J., Seneff S. (1993) A French version of the MIT-ATIS system : portability issues. Actes *3rd European Conference on Speech Communication*, *Eurospeech'93*, Berlin, Allemagne.

[106] Hirschman L. (1998) Language understanding evaluations : lessons learned from MUC and ATIS, Actes *1st Conference on Language Resources and Evaluation*, *LREC'98*. Grenade, Espagne, 117-122. (p. 121-122).

[107] Ferret O., Grau B., Hurault-Planter M., Illouz G., Jacquemin C. (2000) QALC : the question-answering system of LIMSI-CNRS. Actes *Text REtrieval Conference*, *TREC-9*, 235-244 ; Ferret O., Grau B., Hurault-Plantet M., Illouz G., Jacquemin C. (2001) Utilisation des entités nommées et des variantes terminologiques dans un système question - réponse. Actes *TALN'2001*, Tours, France. 153-162.

en évaluation des systèmes de dialogue oral (chapitre 3). Les réflexions du groupe CORAIL et la conception de nos applications s'appuient sur les résultats de ces recherches amont et aval. J'ai donc choisi de leur accorder une place de choix dans ce mémoire. Ce n'est qu'après avoir présenté ces travaux que je m'intéresserai (chapitre 4) aux réalisations de l'équipe. Celles-ci concernent la compréhension automatique de la parole et l'assistance logicielle aux personnes lourdement handicapées.

2. Linguistique de corpus et ingénierie des langues

La langue est une raison humaine qui a ses raisons
Et que l'homme ne connaît pas

Claude Levi-Strauss, *La Pensée sauvage*

La position actuelle des ressources linguistiques en ingénierie des langues a quelque chose de paradoxale. La révolution empiriste qu'a connu le traitement automatique des langues s'est traduite par une redécouverte des corpus. Les ressources linguistiques ont ainsi acquis un rôle déterminant dont témoigne l'existence depuis 1998 d'une conférence internationale (LREC) qui leur est spécifiquement consacrée[1]. Utilisés massivement pour la conception et l'évaluation des systèmes, ces corpus sont cependant considérés avant tout comme de simples données d'apprentissage ou de contrôle. Comme le précisent clairement John Godfrey et Antonio Zampolli[2] :

> *« Their interest, naturally, is in technology and systems that work* […] *(whether scientifically* ***interesting*** *or not) »* [souligné par les auteurs]

D'où ce paradoxe : tout en étant largement reconnues, les ressources linguistiques n'en sont pas moins méconnues. C'est ainsi qu'on assiste à un accroissement continu de la masse des données recueillies pour l'apprentissage sans qu'un effort équivalent ne soit mené en faveur de leur étude linguistique. Alors que les sciences du langage connaissent une mutation profonde avec la (ré)émergence de la linguistique de corpus[3], il y a là une perte de connaissances qui seraient pourtant utiles aux recherches en ingénierie des langues.

C'est en tout cas ce que je vais chercher à montrer dans ce chapitre, en m'intéressant spécifiquement aux études linguistiques de corpus destinés à la communication orale homme - machine. J'en profiterai pour réhabiliter les corpus de dialogue réel homme-homme qui sont souvent délaissés par les approches ingénieriques.

1. LES CORPUS EN CONCEPTION DE SYSTÈMES DE DIALOGUE ORAL

La communication orale homme - machine est un des domaines de l'ingénierie des langues où la nécessité d'une démarche empirique est la plus manifeste. Le concepteur d'un étiqueteur morphosyntaxique (symbolique) peut s'appuyer sur sa propre compétence linguistique pour élaborer un premier prototype aux performances relativement acceptables. Il n'aura ensuite recours à l'étude d'observations réelles que pour améliorer son système. A l'opposé, la conception d'un système de dialogue fondée sur l'intuition apparaît plus difficilement envisageable. Si les membres d'une communauté linguistique partagent plus ou moins la même compétence syntaxique, on observe au contraire des fortes variations de leur comportement dialogique. Même dans le cas d'un dialogue finalisé (restreint à une tâche précise) entre l'homme et la machine, il est difficile d'imaginer intuitivement l'ensemble des formes que peut prendre l'interaction. Cette impuissance concerne aussi bien la structure du dialogue que celle des énoncés produits par l'utilisateur. Nous sommes en effet en présence d'une parole spontanée, enracinée dans l'interaction, qui sera d'autant mieux appréhendée qu'elle sera étudiée à partir d'observations réelles.

---

1 Plus précisément, la conférence *LREC* (*Language Resources and Evaluation Conference*) est consacrée aux ressources linguistiques et à l'évaluation des systèmes d'ingénierie des langues.

2 Godfrey J. J., Zampolli A. (1995) Language resources : overview. *In* Cole R.-A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon. Disponible sur la Toile : http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. 441 :444. Citation page 441.

3 Jean Véronis rappelle ainsi que les corpus intéressent désormais tout autant les linguistes que les chercheurs en ingénierie des langues : « *A l'heure actuelle les corpus […] font partie des outils de base* ***aussi bien des linguistes*** *que des ingénieurs* » [souligné par nous] : J. Véronis (2000). Annotation automatique de corpus : panorama et état de la technique. In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection I²C. Hermès, Paris, France. 235:250.

Aussi est-il fortement recommandé de s'appuyer sur l'observation de données réelles pour concevoir ces systèmes interactifs. Les recommandations du groupe EAGLES sont d'ailleurs très claires sur ce sujet[4]. Plusieurs types de corpus sont utilisables dans le cadre d'une conception par observation. Ils interviendront à des étapes différentes du développement du système.

1.1. Corpus pilotes : définition des besoins par l'analyse d'interactions naturelles

Suivant la terminologie proposée par Jean Caelen et ses collègues, les corpus pilotes regroupent des dialogues réels d'interaction homme-homme destinés à l'étude prospective de la communication homme-machine[5]. Dans cette perspective, les dialogues se focalisent sur l'accomplissement d'une tâche précise. Si on s'intéresse au renseignement aérien, on pourra par exemple enregistrer les conversations reçues par le centre d'appel d'un transporteur[6]. Si un enregistrement en situation réelle n'est pas envisageable, on peut simuler cette dernière en veillant à rester dans le cadre d'une interaction spontanée. On définit alors des scénarii peu contraints qui devront être respectés par de « vrais - faux » interlocuteurs. Aussi lâche soit-elle, cette contrainte influe sur le comportement langagier et dialogique des sujets. Afin de limiter au maximum les biais indésirables, il est nécessaire de définir les scénarii avec le plus grand soin. On rencontre ici une difficulté que l'on retrouvera dans le paragraphe suivant consacré à la simulation des systèmes interactifs.

En règle générale, les corpus pilotes sont utilisés au cours de la phase préliminaire de définition du système, où leur utilisation est à rapprocher de l'étape d'analyse des besoins du génie logiciel. En particulier, ils constituent un outil apprécié pour comprendre et circonscrire l'univers « naturel » de la tâche. Leur importance est ainsi résumée par les experts du groupe EAGLES[7] :

> *« RECOMMANDATION — Where possible, use human-human dialogue data to build an understanding of the domain and its component tasks »*

Ces ressources exploratoires définissent donc — « pilotent » — la feuille de route qui guidera la conception du système.

Pour l'ingénierie des langues, l'utilisation des corpus pilotes se limite généralement à cette caractérisation de la tâche (cf infra §1.5). Les corpus pilotes attestent pourtant d'usages langagiers qui intéressent la conception des systèmes. Dybkjaer et Bersen, qui se sont intéressés de longue date à l'utilisabilité des systèmes interactifs, rappellent ainsi l'importance des corpus pilotes pour la prise en compte de l'utilisateur final[8]. Jean Caelen rappelle à juste titre que leur étude est incontournable dans le cas des prototypes de laboratoire dont on ne connaît pas encore la signification d'usage[9].

La portée des analyses d'usage sur corpus pilotes peut même dépasser la seule problématique du dialogue homme-machine. Je reviendrai sur ce point en fin de chapitre.

1.2. Simulation par magicien d'Oz : adaptation du comportement face à la machine

Aussi conviviaux soient-ils, les systèmes de dialogue induisent une modification du comportement langagier de leurs utilisateurs. Plusieurs études ont ainsi montré que les êtres humains adaptaient leur expression langagière et leurs stratégies de dialogue en fonction de la nature (personne ou

---

4 Fraser N. (1997) Assessment of interactive systems. *In* Gibbon D., Moore R., Winski R. (Eds.). (1997) Handbook of standards and resources for spoken language systems. Mouton de Gruyter, Berlin, Allemagne. 564-615.

5 Caelen J., Zeiliger J., Bessac M., Siroux J., Perennou G. (1997) op. cit.

6 Il conviendra de respecter les règles déontologiques en usage. On veillera ainsi à anonymer les transactions et à recueillir l'accord des intéressés. En dehors de la législation sur le droit à l'image et la loi « Informatique et Libertés », il existe dans le droit français un certain vide juridique en matière de constitution et de diffusion de corpus. Il serait souhaitable que l'État clarifie cette situation afin de limiter le risque juridique qui pèse actuellement sur les fournisseurs de corpus.

7 Fraser N. (1997) op. cit. (citation p. 580).

8 Dybkjaer L., Bersen N.-O. (2000) Usability issues in spoken dialogue systems. *Natural Language Engineering*, 6 (3-4), 243-271. (paragraphe II, p. 247).

9 Caelen J., Zeiliger J., Bessac M., Siroux J., Perennou G. (1997) op. cit.

machine) de leur interlocuteur[10]. Dans un article abordant cette question, Jean-Marie Pierrel fait remarquer que l'utilisateur essaie de s'adapter aux scénarii supposés connus par la machine, de même qu'il s'efforce d'être le plus complet et le plus clair possible dans ses requêtes[11]. Au final, il observe que le dialogue entre l'utilisateur et la machine est sensiblement moins interactif qu'une interaction entre deux interlocuteurs humains sur la même tâche.

Cette adaptation répond au souci qu'a l'utilisateur d'aider des systèmes qui restent pour l'instant incapables de réussir le célèbre test de Turing[12]. Sans préjuger des avancées futures du dialogue homme - machine, on peut se demander si nous sommes uniquement en présence d'une adaptation palliative. Les études de Marie-Annick Morel et ses collègues sur les corpus CIO et SNCF ont ainsi observé des variations d'usages qui ne reposaient que sur les croyances du locuteur[13]. Alors qu'ils se trouvaient en présence d'un système simulé par un être humain, leur comportement était différent s'ils étaient préalablement informés de la simulation. Dans les deux cas, le comportement du système simulé était pourtant identique. Nous sommes certainement en présence de considérations d'ordre sociologiques ou psychologiques qui conduisent les êtres humains à refuser de dialoguer avec les systèmes artificiels comme avec leurs semblables[14].

Du fait de cette modification irrépressible du comportement des utilisateurs, les corpus pilotes ne peuvent constituer qu'une idéalisation du dialogue homme – machine. Il importe donc de disposer également de corpus de dialogue homme - machine sur lesquels on étudiera ces adaptations d'usage.

Comme dans le cas des corpus pilotes, ces dialogues peuvent être réels ou simulés. Dans ce dernier cas, on utilise la technique dite du magicien d'Oz[15]. Celle-ci consiste à simuler le système par une maquette contrôlée par un opérateur humain (le « compère »). La simulation est généralement réalisée à l'insu de l'utilisateur. On parle alors de vrai magicien d'Oz, par opposition au paradigme de faux magicien d'Oz où le sujet est averti du caractère factice du système. Dans une phase de prototypage, cette technique est très utile puisqu'on ne dispose pas encore de système opérationnel. Son intérêt dépasse la simple simulation du système à concevoir, puisqu'elle permet de tester différentes stratégies de dialogue ou d'évaluer l'influence des contextes d'utilisation sans avoir à implémenter de système.

En règle générale, les expériences de magicien d'Oz mettent en jeu des maquettes simulant le comportement d'un système de dialogue optimal. Les corpus recueillis permettent ainsi une étude précise des usages langagiers attendus. La technique du magicien d'Oz a ainsi largement démontré son intérêt pour le développement, mais aussi la validation, des systèmes interactifs.

La technique du magicien d'Oz demande cependant un maquettage fin pour arriver à une simulation réaliste n'induisant pas de modification inopportune du comportement des sujets. La simulation, qui concerne aussi bien les entrées / sorties du système (reconnaissance, compréhension et génération) que la gestion du dialogue, ne peut être atteinte que par la satisfaction de trois contraintes :

a) elle doit répondre à la spécification du système attendu. Dans le cas de systèmes innovants, le recours préalable à une analyse des usages sur corpus pilote est souhaitable. Cette étude amont

---

10 Hauptman A., Rudnicky A. (1988) Talking to computers : an empirical investigation. *International Journal of Man-Machine Studies*, 28. 583-604 ; Morel M.A. (Ed.) (1989) Analyse linguistique d'un corpus ; 2° corpus : centre d'information et d'orientation de l'Université Paris V. Publications de la Sorbonne Nouvelle, Paris, France ; Spérandio J.-C., Létang-Figeac C. (1986) Simulation expérimentale de dialogues oraux en communication homme - machine. Rapport final GRECO Communication Parlée. CNRS, Paris, France.

11 Pierrel J.-M. (1988) Dialogue homme - machine en langage naturel écrit et oral. Actes *1ères journées nationales du PRC Communication Homme - machine*, Ec2 Editions, Paris, France. 152-182.

12 Turing A. M. (1950) Computing machinery and intelligence. *Mind*. 59, 236.

13 Morel M.A. (Ed.) (1989) op. cit.

14 Button G. (1990) Going up a blind alley : conflating conversation analysis and computational modelling. In Luff P., Gilbert G., Frohlich D. (Eds.) Computers and conversation. Academic Press, London, Royaume-Uni. 67-90 ; cité par : Fraser N. (1997) Assessment of interactive systems. In Gibbon D., Moore D., Winski R. (Eds). *Handbook of standard and resources for spoken language systems*. Mouton de Gruyter, Berlin, Allemagne. 564-651 (partie concernant le magicien d'Oz : 581-591)

15 Fraser N., Gilbert G. (1991) Simulating speech systems. *Computer Speech and Language*, 5. 81-99.

permet de définir l'univers de la tâche et de bâtir des scénarii d'interaction réalistes,

b) elle doit modéliser les limitations techniques (prévisibles) du système, faute de quoi on ne peut étudier les capacités d'adaptation palliatives de l'utilisateur. La simulation réaliste des erreurs est cependant assez difficile, et doit répondre là encore à des spécifications précises,

c) enfin, il faut s'assurer que la charge cognitive que représente le contrôle de la maquette de simulation n'est pas trop importante pour le compère.

On le voit, la mise en place d'une expérience de magicien d'Oz est toujours délicate. Afin de limiter les biais de conception de la maquette, Marc Guyomard et Jacques Siroux proposent une démarche incrémentale de réalisation en deux étapes[16]. Il n'en reste pas moins très difficile de réaliser une simulation complète qui rend compte de toutes les situations d'interaction. Comme le précisent Fraser et Gilbert[17] :

> « *Very few* [Wizard of Oz] *experiments have attempted to simulate all the components of a speech dialogue system* »

Par sa lourdeur de mise en œuvre, la simulation interdit par ailleurs l'acquisition de corpus d'observations d'envergure[18]. Les corpus simulés sont donc surtout utilisés pour affiner l'analyse du problème qui a été définie par l'étude de corpus pilotes. La technique du magicien d'Oz autorise ainsi un maquettage fonctionnel suffisamment précis pour se lancer dans la conception du système proprement dit. Cette dernière étape fait alors usage de corpus réels de dialogue homme - machine.

1.3. Corpus de dialogues homme - machine réels : conception par bootstrap

Les corpus de dialogues homme - machine réels sont au centre des méthodologies de conception incrémentale qui se sont généralisées en CHM. Cette démarche consiste à réaliser rapidement un premier prototype imparfait qui sera ensuite amélioré par des cycles de conception / validation conduisant à chaque étape à la réalisation d'un nouveau prototype. L'intérêt de cette approche réside dans l'étape de validation, qui est menée en conditions réelles d'utilisation. L'analyse des erreurs qui sont relevées au cours de ces sessions d'utilisation guide la réalisation de la version suivante du prototype. On parle ainsi d'approche « *system-in-the-loop* » ou de conception par *bootstrap*[19].

Cette conception itérative, qui intègre l'utilisateur tout au long du développement du système, a largement fait ses preuves en génie logiciel (cycle de vie en spirale[20]). Elle constitue une démarche efficace pour construire des systèmes interactifs opérationnels. Il importe cependant que l'univers de la tâche ait été préalablement circonscrit avec attention et que l'on ait une idée précise des attentes des utilisateurs. Dans le cas contraire, les capacités d'adaptation palliative d'utilisateurs cherchant à se conformer aux limitations du prototype sont susceptibles de masquer des problèmes d'utilisabilité majeurs. Seules des expérimentations en situations réelles mettant en jeu des

---

16 Guyomard M., Siroux J. (1987) Experimentation in the specification of an oral dialogue. *In* Niemann H., Lang M., Sagerer G. (Eds.) Recent advances in speech understanding and dialog systems. *Computer and System Sciences*. 46, Springer Verlag, Berlin, Allemagne. 497-501.

17 Fraser N., Gilbert G. (1991) Effects of system voice quality on user utterances in speech dialogue systems. Actes *2nd European Conference on Speech Communication and Technology, Eurospeech'91*, Gènes, Italie, 57-60.

18 C'est d'ailleurs le principal reproche que les participants du projet européen ESPRIT Mask (renseignement ferroviaire) adressent à la technique du magicien d'Oz : Life A., Salter I., Temem J.N., Dartigues H., Guidon A., Rosset S., Bennacef S., Lamel L. (1996) Data collection for the MASK kiosk : Woz versus prototype system. Actes *4th International Conference on Spoken Language Processing*, *ICSLP'1996*, Philadelphie, PA, Etats-Unis. 1672-1675. Disponible sur la Toile : http://www.asel.udel.edu/icslp/cdrom/vol3/658/a658.pdf

19 Lamel L., Rosset S., Bennacef S., Bonneau-Maynard H., Devillers L., Gauvain J.-L. (1995) Development of spoken language corpora for travel information. Actes *4th European Conference on Speech Communication and Technology*, *Eurospeech'95*. Madrid, Espagne. 1961-1964.

20 Boehm B. W. (1988) A spirale model of software development and enhancement. *IEEE Computer*, 21(5). 61-72 ; Boehm B.W., Gray T.E., Seewaldt T. (1984) Prototyping versus specifying : a multi-project experiment. *IEEE Transactions on software engineering*, 10(3). 290-303.

utilisateurs non recrutés pourraient alors révéler ces limitations. Compte tenu de leur lourdeur, ces évaluations *in situ* sont malheureusement limitées[21].

Le statut de ce type de ressource linguistique est donc clair : il s'agit de corpus de développement destinés aux concepteurs des systèmes pour une utilisation purement ingénierique. Du fait de leur caractère idiosyncrasique (un corpus = un système), leur utilisation est a priori très restreinte. Il me semble cependant que ces ressources contiennent des observations susceptibles d'intéresser l'ensemble de la communauté scientifique. Par exemple, on peut envisager d'étudier sur ces données le comportement des utilisateurs face à une situation d'échec. Plus largement, le dialogue est une co-construction entre deux interlocuteurs. Ces corpus peuvent donc constituer un matériel original pour observer cette construction, jusque dans ses erreurs et malentendus.

1.4. Corpus et conception de systèmes interactifs

Les types de corpus que nous venons d'étudier jouent ainsi un rôle complémentaire dans la conception des systèmes interactifs. Si on synthétise les recommandations du groupe EAGLES[22], on constate que cette conception devrait suivre idéalement un processus de développement qui rappelle le cycle en " b " défini par le génie logiciel[23]. La figure 2.1 (page suivante) présente les différentes étapes qui constituent ce cycle de vie. Nous allons les étudier brièvement en faisant le lien avec les trois types de corpus caractérisés plus haut.

Analyse des besoins — Une première étape d'analyse des besoins consiste à délimiter l'univers de la tâche et à déterminer les usages langagiers et dialogiques qui devront être modélisés par le système. Afin de prendre en considération les besoins réels des utilisateurs, cette analyse doit être menée sur des situations interactives écologiques, c'est-à-dire dans le cadre d'interactions humaines réelles correspondant à la tâche. On fera donc appel ici à des corpus pilotes.

En génie logiciel, l'analyse et la spécification des besoins servent également à définir un cahier de recettes sur lequel s'appuie la validation du système. Cette pratique peut et doit être reprise en ingénierie des langues. Comme le précisent Jean Caelen, Jérôme Zeiliger, Jacques Siroux et leurs collègues[24] :

> *« Dès l'étape* [de spécification des besoins] *on peut évaluer si le cahier des charges est cohérent et répond bien au besoin. Pour cela, encore faut-il avoir étudié les usages et ciblé correctement les besoins. En dialogue homme - machine, cela doit être le rôle du(es) corpus pilote(s) ».*

Suivant une démarche reprise du modèle en V du génie logiciel, l'analyse des besoins va servir de guide à l'étape suivante d'analyse fonctionnelle.

---

21 Certains projets ont donné lieu à de telles expérimentations. Dans le cas du projet européen ARISE, qui visait la réalisation de systèmes de renseignement ferroviaire, le système du LIMSI a par exemple été testé sur des appels réels retenus par le SNCF : den Os E., Boves L., Lamel L., Baggia P. (1999). Overview of the ARISE project, Actes *6th European Conference on Speech Communication and Technology*, *Eurospeech'99*, Budapest, Hongrie. 1527-1530 ; Rosset S. (2000) Stratégies et gestionnaire de dialogue pour des systèmes d'interrogation de bases de données à reconnaissance vocale. Doctorat Université Paris XI, Orsay, France. Publié comme rapport de recherche 2001-18 du LIMSI-CNRS, Orsay, France. septembre 2001.
De même, le système MASK du LIMSI (borne de renseignement touristique) a donné lieu a une expérimentation in situ. Les sujets de l'expérimentation conservaient leur libre arbitre puisqu'ils étaient simplement invités à utiliser un système dont ils étaient averti du caractère artificiel.

22 Gibbon D., Moore R., Winski R. (Eds.). (1997) *op. cit*.

23 Le cycle en " b " n'est en fait qu'une amélioration du cycle de développement en V utilisée en génie logiciel (le caractère ascendant du cycle en V n'est pas représenté par la figure 2.1.) : Jacobson I. (1993) Le génie logiciel orienté objet : une approche fondée sur les cas d'utilisation. ACM Press, Addison Wesley.

24 Caelen J., Zeiliger J., Bessac M., Siroux J., Perennou G. (1997) *op. cit.* (citation page 216).

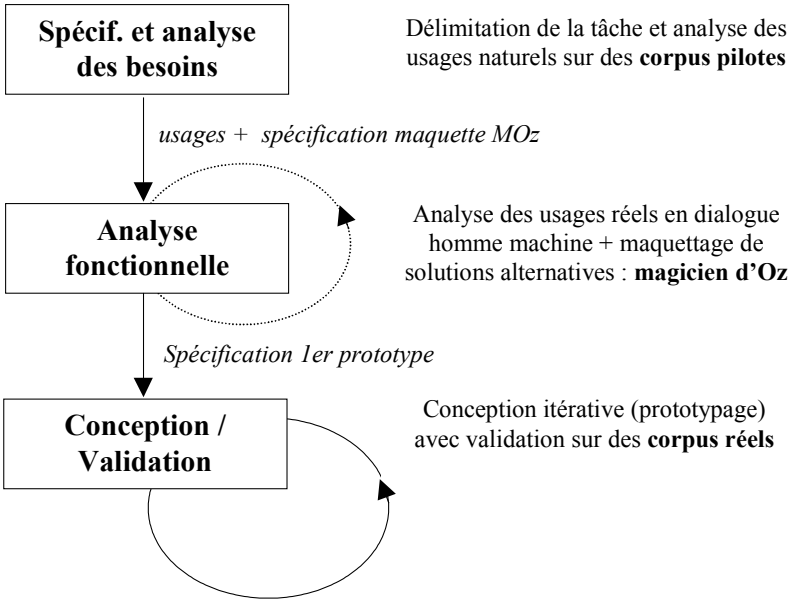

Figure 2.1 *— Conception des systèmes interactifs suivant un cycle de vie en " b ".*

Analyse fonctionnelle — L'objectif de cette étape est de fournir une spécification logicielle précise du système interactif. Cette spécification repose sur l'analyse de corpus simulés par la technique du magicien d'Oz. La maquette de simulation qui est utilisée doit répondre aux spécifications issues de l'analyse des besoins précédente.

La simulation permet tout d'abord de préciser l'analyse des besoins précédente en y intégrant certaines modifications d'usage inhérentes au dialogue homme - machine. Ensuite, elle autorise l'étude de solutions alternatives de conception sans avoir à se lancer dans le développement de plusieurs prototypes. Par exemple, différentes stratégies de dialogues peuvent être simulées au cours de l'expérimentation.

Les processus de conception incrémentale qui se sont développés en génie logiciel ont eu tendance à remplacer ce maquettage exploratoire par des cycles de prototypages supplémentaires. La recherche scientifique comportant une importante part d'inconnue, il semble encore pertinent de recourir à une évaluation rapide de grandes pistes de recherche à l'aide de la simulation.

Suivant le modèle en V du génie logiciel, l'étape d'analyse sert elle aussi à préparer la validation du système. Les corpus simulés peuvent donc être utilisés, au même titre que les corpus pilotes, pour préparer l'évaluation du système. A titre d'exemple, les corpus de magicien d'Oz du projet MASK (renseignement ferroviaire) ont servi avant tout à la définition des tests de l'interface utilisateur[25].

Conception itérative — Les spécifications fonctionnelles fournies par la simulation précédente permettent de guider le développement du système proprement dit. On emploie alors une démarche itérative par prototypages successifs. La validation des prototypes repose sur des expérimentations aussi réelles que possible. L'analyse de ces sessions d'utilisation va faire apparaître des limitations auxquelles devront répondre les prototypes suivants. La résolution des erreurs ne doit pas perdre de vue les enseignements des étapes précédentes d'analyse. Il est donc souhaitable que la validation concerne également la satisfaction des usages caractérisés sur corpus pilotes ou simulés.

1.5. Corpus pilotes pour une conception plus linguistique

Le processus de développement que je viens d'esquisser ne se retrouve malheureusement que très partiellement dans les pratiques quotidiennes du dialogue oral homme – machine. Dans les années 1980, le dialogue oral homme - machine accordait une attention significative à l'analyse de corpus

[25] Life A., Salter I., Temem J.N., Dartigues H., Guidon A., Rosset S., Bennacef S., Lamel L. (1996) *op. cit.*.

réels ou de magicien d'Oz[26]. A l'opposé, la révolution empiriste qu'a connu l'ingénierie des langues s'est traduite par un recours limité à ces études amont. On peut d'ailleurs s'interroger sur l'influence qu'ont eu les campagnes d'évaluation à caractère ingénierique sur cette évolution (cf. chap. 3).

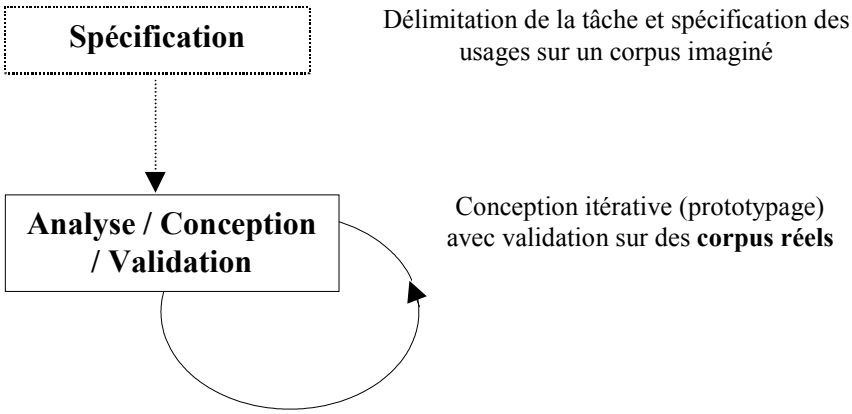

***Figure 2.2** — Conception des systèmes de dialogue oral homme - machine suivant un cycle de vie exclusivement itératif*

La conception des systèmes interactifs s'apparente ainsi le plus souvent à un cycle de vie totalement incrémental d'où a disparu l'étape préalable d'analyse des usages (figure 2.2.). La délimitation de la tâche peut provenir directement des spécifications de la campagne de test, de même que le premier prototype de système peut résulter de l'analyse d'un corpus imaginé par le concepteur lui-même. Lorsqu'elle décrit la méthodologie itérative de conception du LIMSI, Sophie Rosset ne parle pas de collecte de données pour la spécification du système[27]. Le premier prototype est directement élaboré à partir d'une grammaire minimale de compréhension et d'un module de dialogue préliminaire qui sont fondés sur l'intuition de leur concepteur.

Le recours à l'observation n'intervient qu'ultérieurement, lorsqu'on procède à la validation des prototypes successifs du système (fichiers de *log* ou corpus de test d'une campagne d'évaluation). Il me semble que cette conception centrée sur des corpus de développement comporte certaines limitations importantes :

- à moyen terme, l'adaptation palliative des utilisateurs aux insuffisances du système peut masquer des problèmes significatifs d'utilisabilité du système,
- à long terme, ces améliorations successives à partir de jeux d'essais bien balisés risquent de masquer des questions de fond qu'aurait pu soulever la confrontation des usages réels aux usages attendus.

Dans la perspective d'une ingénierie des langues centrée sur l'utilisateur, il me semble au contraire important d'acquérir des connaissances solides sur les besoins et usages langagiers qui sont sous-jacents au dialogue homme - machine. C'est pourquoi une part significative de mes travaux est consacrée à l'étude de corpus d'études, et plus particulièrement de corpus pilotes. Je présenterai dans ce chapitre plusieurs résultats qui suggèrent l'intérêt de cette démarche pour la communication homme - machine mais aussi plus généralement pour l'ingénierie des langues (§ 3). Cette démarche pose la question de la disponibilité de corpus francophones de dialogue oral, sur laquelle je vais tout d'abord revenir.

---

[26] Voir par exemple les travaux du GDR-PRC Communication Homme - machine. Citons par exemple les corpus recueillis par Marie-Annick Morel et ses collègues (corpus CIO, SNCF, Air France), dont certains ont ensuite été repris et étudiés dans le cadre du projet DALI du GDR-PRC (rapport final du projet DALI : http://www-geod.imag.fr/pages_html/projets/DALI.html)

[27] Rosset S. (2000) Stratégies et gestionnaire de dialogue pour les systèmes d'interrogation de bases de données à reconnaissance vocale. Thèse de Doctorat. Université Paris XI, Orsay, France. 15 décembre 2000. (page 97-101).

2. POUR UNE LIBRE DIFFUSION DES CORPUS DE FRANÇAIS PARLE

2.1. Le retard du français

A l'heure de la généralisation de l'outil informatique et de l'accès à de grandes banques de connaissances par l'intermédiaire d'Internet, la place d'une langue dans notre société mondialisée dépend de ses possibilités d'utilisation dans les nouvelles technologies d'information et de communication (NTIC). L'ingénierie des langues se doit donc de proposer pour chaque idiome un ensemble de technologies langagières permettant la recherche et la restitution d'informations sous forme écrite, orale ou multimédia.

2.1.1. Des corpus spécifiques pour des technologies adaptées à chaque idiome

Cette exigence pose bien entendu le problème de la généricité idiomatique des méthodes utilisées en ingénierie des langues[28]. Par delà cette question essentielle, le développement des technologies langagières sur un idiome particulier impose également la mise en place de ressources linguistiques significatives dans cette langue. L'exemple de l'*American National Corpus*[29] illustre la criticité de cette question. Alors que l'on dispose avec le *British National Corpus* (*BNC*) d'une ressource considérable sur l'anglais britannique[30], le LDC (*Linguistic Data Consortium*) américain s'est en effet engagé dans le développement d'un corpus équivalent pour l'anglais américain. Lorsqu'on connaît la proximité de ces deux idiomes, on saisit la nécessité de la mise en place de ressources linguistiques équivalentes pour le français.

2.1.2. Ressources linguistiques francophones : quelques données quantitatives

Avec des ressources pionnières telles que le *Trésor de la Langue Française*[31] pour l'écrit, ou BDSONS[32] pour l'oral, le français était certainement au début des années 1980 une des langues les mieux outillées en corpus. Malheureusement, le développement de l'ingénierie des langues ne s'est pas accompagné d'un accroissement comparable des ressources linguistiques francophones. Au contraire, le retard pris par le français dans ce domaine est désormais considérable :

- On dispose à l'heure actuelle pour l'anglais écrit d'un corpus annoté grammaticalement de 100 millions de mots (*BNC*) alors que la communauté francophone ne dispose même pas d'un corpus équivalent d'un million de mots[33],

---

[28] En dialogue oral homme - machine, cette question n'est généralement abordée qu'au cas par cas lors du portage d'un système d'une langue à une autre (Bonneau-Maynard H., Gauvain J.-L., Goodine D., Lamel L., Polifroni J., Seneff S. 1993. A French version of the MIT-ATIS system : portability issues. *Eurospeech'93*, Berlin, Allemagne). Or, on peut s'interroger plus globalement sur le caractère anglo-centré de la plupart des théories et méthodes utilisées en ingénierie des langues. L'anglais est en effet une langue très particulière dont la morphologie très pauvre est compensée par de fortes contraintes d'ordonnancement linéaire. On sait que cette rigidité est bien adaptée aux modélisations markoviennes du TAL probabiliste. A l'opposé, ces méthodes créées à l'origine pour l'anglais s'appliquent malaisément à des langues à ordre libre comme le coréen, le russe ou le finnois (Covington M.A., 1990, A dependency parser for variable-order languages, rapport de recherche AI-1990-01, University of Georgia, Etats-Unis ; Koo M. W. *et al*., 1995, KT-STS : A speech translation system for hotel reservation and a continuous speech recognition system for speech translation, *Eurospeech'95*, Madrid, Espagne, 1227:1231).

[29] Ide N., Macleod C. (2001). The American National Corpus : a standardized resource for American English. Actes *Corpus Linguistics'2001*, Lancaster, Royaume-Uni, 274 :280

[30] Leech G., Garside R., Bryant M. (1994). CLAWS4 : The tagging of the British National Corpus. Actes *14th International Conference on Computational Linguistics*, *COLING'1994*, Kyoto, Japon, 622 624.

[31] Constitué à partir des années 1960, ce corpus essentiellement littéraire est également structuré sous forme de dictionnaire. Il est diffusé librement par l'ATILF (ex-INALF) en version informatisée (TLFi) : Bernard P., Dendien J., Lecomte J., Pierrel J.-M. (2002) Un ensemble de ressources informatisées et intégrées pour l'étude du français : FRANTEXT, TLFi, Dictionnaires de l'Académie et logiciel Stella. Actes *TALN'2002*, Nancy, France. 3-36 ; Bernard P., Bernet C., Dendien J., Pierrel J.-M., Souvay G., Tucsnak Z. (2001) Ressources linguistiques informatisées de l'ATILF. Actes *TALN'2001*, Tours, France. vol 1, 333-338.

[32] Carré R., Descout R., Eskénazi M., Mariani J., Rossi M. (1984). The French language database : defining, planning and recording a large database. Actes *1984 International Conference on Acoustics, Speech and Signal Processing*, *ICASSP'1984*., San Diego, CA. Vol. 3 42-10.1 - 42.10-4.

[33] Véronis J. (2000). Annotation automatique de corpus : panorama et état de la technique. In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection I²C. Hermès, Paris, France. 235 :250. (données citées page 112).

- Dans le domaine des corpus arborés en structures syntaxiques (*treebanks*), l'anglais dispose depuis une dizaine d'années de plusieurs corpus[34] (*Penn Treebank*, corpus *IBM/Lancaster*) regroupant chacun plus de 3 millions de mots. Le laboratoire TALANA travaille depuis plusieurs années sur la constitution d'un corpus arboré francophone[35] de taille sensiblement inférieure. Cette ressource est enfin diffusée (librement pour la recherche académique) depuis cette année.
- Le *BNC* comporte des corpus oraux regroupant plus de 10 millions de mots de parole transcrite. Pour le français, seuls trois corpus oraux (CORPAIX, ELILAP et VALIBEL) présentent une taille atteignant le million de mots. Leur statut est en outre relativement précaire. Le corpus CORPAIX[36], collecté depuis de nombreuses années par le GARS à Aix-en-Provence, n'est que très partiellement informatisé. Il n'est de toute manière pas distribué. Il en va de même pour le corpus VALIBEL[37] de l'Université de Louvain-la-Neuve (Belgique) qui forme la plus grande ressource orale en francophonie (3 900 000 mots pour 375 heures d'enregistrement). A l'opposé, le corpus ELILAP de l'université de Leuven (Belgique) est interrogeable librement sous format électronique[38]. Comme CORPAIX, il ne regroupe malheureusement que des monologues ou des interviews peu interactifs. Son utilisation en dialogue oral homme - machine est donc limitée[39]. Dans cette perspective, les seules ressources largement diffusées sont les corpus pilotes ou de magicien d'Oz (corpus Air France[40], CIO, SNCF) qui avaient été collectés au cours des années 1980 dans le cadre du GDR-PRC Communication Homme - Machine. Ils regroupent chacun moins de 50 000 mots. Cette taille devrait être dépassée, pour ce qui concerne le dialogue oral proprement dit, par les corpus *Français de Référence*[41] et, surtout, C-ORAL-ROM[42] sur lesquels travaille actuellement le laboratoire DELIC. Ce dernier corpus, qui devrait avoisiner la centaine de milliers de mots dans sa partie interaction orale, ne sera toutefois pas distribué librement.

A ma connaissance, le plus grand corpus francophone annoté actuellement distribué a été réalisé dans le cadre du projet MULTEXT (corpus JOC français). Il est commercialisé par l'ELDA et regroupe 200 000 mots environ.

34 Marcus M., Santorini B., Marcinkiewicz M. (1993). Building a large annotated corpus of English : the Penn Treebank. *Computational Linguistics*, 19(2), 313-330 ; Leech G., Garside R. (1991) Running a grammar factory : the production of syntactically analysed corpora or " treebanks ". In Johansson S., Strenström A.-B. (Eds.) *English computer corpora : selected papers and research guide*, Mouton de Gruyter, Berlin, Allemagne. 15-32.

35 Abeillé A., Clément L., Reyes R. (1998). TALANA annotated corpus : the first results. Actes *1st Conference on Linguistic Resources and Evaluation*, *LREC'1998*, Grenade, Espagne, 992-999 ; Abeillé A. , Clément L., Kinyon A. (2000) Bulding a treebank for French. Actes *2nd Conference on Linguistic Resources and Evaluation*, *LREC'2000*, Athens, Greece, 87-94.

36 Ce corpus regroupe environ un million de mots. Pour un aperçu du contenu de ce corpus non diffusé : Blanche-Benveniste C., Rouget C., Sabio F. (2002) Choix de textes de français parlé : 36 extraits. Honoré Champion, Paris.

37 Dister A. (2002) Normalisation de corpus oraux retranscrits : jusqu'à quel point ? Actes des *2ème journées de la Linguistique de Corpus*, Lorient, France. p. 15 (résumé). Toile : http://valibel.fltr.ucl.ac.be/val-banque.html.

38 Ce corpus, qui regroupe environ 900 000 mots transcrits, reprend une partie du corpus d'Orléans enregistré entre 1968 et 1971 (http://bach.arts.kuleuven.ac.be/elicop/projetELILAP.htm). Il est désormais intégré dans le corpus ELICOP : http://bach.arts.kuleuven.ac.be/elicop/

39 Kerbrat-Orecchioni a montré combien le degré d'interactivité influe sur les productions orales : Kerbrat-Orecchioni C. (1999) L'oral dans l'interaction : une liberté surveillée. *Revue Française de Linguistique Appliquée*, 4(2), 41-55.

40 Ce corpus pilote de réservation et renseignement aérien a été enregistré entre 1988 et 1990 par M.-A. Morel et D. Delomier (Université de la Sorbonne Nouvelle, Paris 3) puis révisé par Pierre Nerzic (IRISA, Lannion) dans le cadre du projet DALI. Disponible sur la Toile : http://www.inalf.fr/ananas/site/htm/AirFrance.html

41 Ce corpus de référence de français parlé devrait regrouper à terme 450 000 mots. Son objectif est d'offrir une description sociolinguistique d'ensemble du français parlé. Les concepteurs du corpus ont recherché des situations privilégiant les productions orales assez longues pour faciliter l'observation de phénomènes syntaxiques. Ce corpus comporte donc peu de dialogues oraux. Il devrait être commercialisé par l'ELDA : DELIC (2002) Le corpus de référence de français parlé. Actes *2èmes journées de la linguistique de corpus*, Lorient, France, p. 41 (résumé). Voir aussi http://www.up.univ-mrs.fr/veronis/Atala/jecorpus/Bilger.htm.

42 Le projet européen C-ORAL-ROM vise la constitution d'un corpus oral représentatif dans les principales langues romanes. Chaque corpus national regroupera un échantillonnage de situation énonciatives, parmi lesquelles le dialogue oral informel (50 000 mots à 100 000 mots par idiomes). Ce corpus sera commercialisé par l'ELDA en 2004 : Cresti E. *et al.* (2002) The C-ORAL-ROM project. New methods for spoken language archives in a multilingual romance corpus. Actes *3rd International Conference on Language Resources and Evaluation*. *LREC'2002*. Las Palmas de Gran Canaria. Espagne. vol. I, 2-9.

- Enfin, il n'existe pas de corpus oral significatif annoté en parties du discours, contrairement au BNC anglais[43].

Si la prédominance de l'anglais en matière de ressources linguistiques est réelle, on observe en outre que l'allemand, le néerlandais, le suédois ou encore le japonais ont bénéficié d'efforts qui leur permettent de surpasser très largement la taille des corpus francophones actuellement disponibles.

Deux causes semblent à l'origine de cette situation inquiétante. La première réside dans l'absence, jusqu'à une date récente, d'une politique cohérente des ministères français de la recherche et de la francophonie[44] en faveur des ressources linguistiques. Il est symptomatique que les principales ressources francophones actuellement disponibles proviennent de recherches menées au Canada (corpus bilingue aligné HANSARD par exemple), en Belgique (corpus oraux ELILAP et VALIBEL déjà cités), voire par des département de français langue étrangère. Le ministère de la recherche semble enfin avoir pris conscience de l'importance de ce problème. Le récent appel d'offre TECHNOLANGUE comprend ainsi un volet ressources linguistiques qui devrait favoriser non seulement la constitution de corpus francophones, mais surtout leur diffusion.

De nombreuses ressources d'envergure existent pourtant au sein des laboratoires français. En l'absence de politique nationale valorisant leur diffusion, ces laboratoires ont adopté une attitude de repli et ont conservé en interne des données dont le recueil leur avait demandé des efforts de longue haleine. Comme le fait remarquer Laurent Romary[45], en diffusant ces corpus,

> *« [ces équipes] ont l'impression de se faire piller ou, tout simplement, de perdre une partie de ce qui fait leur connaissance scientifique. »*

La diffusion des « richesses » (corpus ou outils) d'un laboratoire peut pourtant s'avérer gratifiante pour ce dernier. Laurent Romary poursuit ainsi[46] :

> *« Ce dernier argument s'avère dans les faits trompeur : les équipes qui se sont engagées dans la voie de la diffusion large de leurs ressources et des méthodologies associées ont bénéficié d'un regain de renommée non négligeable ».*

C'est pourquoi j'ai engagé le groupe CORAIL dans une politique systématique de diffusion de nos ressources. Mon premier souci est tout d'abord de faire vivre au maximum les données que nous avons recueillies. Celles-ci peuvent en effet rencontrer un intérêt en dehors des problématiques qui ont présidé à leur constitution. Mais plus généralement, mon ambition est de contribuer au développement des ressources linguistiques en français. Le groupe CORAIL s'est ainsi lancé dans un programme de constitution de corpus de dialogue oral (programme *Parole Publique*) destinés à être diffusés librement. Ce type d'activité est peu valorisé scientifiquement. Il n'en est pas moins essentiel au développement de l'ingénierie des langues. Il me semble donc important de présenter ici cette initiative.

2.2. Parole Publique : une initiative pour la libre diffusion de corpus de dialogue oral

2.2.1. Objectifs

Le programme *Parole Publique*[47] s'intègre à mes activités en Communication Homme-Machine. Plus précisément, il s'intéresse à la constitution de corpus d'études (corpus pilotes ou de magicien d'Oz) utiles aux recherches en dialogue oral homme-machine. Il ambitionne à terme la réalisation

[43] Valli A. et Véronis J. (1999) *op. cit.*

[44] Les actions menées par l'AUF (*Agence Universitaire de la Francophonie*, ex AUPELF-UREF) entre 1995 et 2000 furent ainsi loin d'être à la mesure des besoins des chercheurs francophones. Depuis cette date, l'AUF a abandonné toute politique de soutien à la recherche en ingénierie des langues.

[45] Romary L. (2000) Outils d'accès à des ressources linguistiques. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 193-212. Citation page 194.

[46] *Op. cit.* Citation page 194-195.

[47] Site du projet *Parole Publique* sur la Toile : http://www.univ-ubs.fr/valoria/antoine/parole_publique

d'une ressource échantillonnée rendant compte de contextes interactifs variés. Cette banque de corpus serait ainsi représentative de la diversité du dialogue oral finalisé[48].

Pour le programme *Parole Publique*, cette diversité concerne avant tout la tâche et le domaine d'application au sein desquels l'interaction est circonscrite. Il s'agit d'une dimension essentielle pour le dialogue homme-machine. La constitution d'une banque de corpus couvrant différents domaines d'application pourrait en effet lever certaines interrogations sur la généricité des systèmes interactifs[49]. Les financements dont je dispose à l'heure actuelle nous permettent de nous intéresser aux domaines d'applications suivants :

- renseignement touristique,
- réservation hôtelière,
- renseignement administratif,
- portail vocal entreprise,
- accueil standard téléphonique.

Cette couverture thématique peut cependant être étendue à l'avenir.

L'objectif du programme *Parole Publique* n'est pas d'atteindre un échantillonnage sociolinguistique équilibré des locuteurs enregistrés. L'adaptation des systèmes au type d'usager constitue cependant une forme de généricité essentielle à la réalisation d'applications conviviales. Pour l'heure, *Parole Publique* aborde cette problématique uniquement du point de vue de l'âge des locuteurs (adolescents, adultes et personnes âgées). A terme, j'espère que nous disposerons d'une collection de corpus suffisamment diversifiée pour autoriser une analyse différentielle des usages langagiers vis à vis de cette variable.

Précisons enfin que si le mode d'interaction orale est privilégié dans ce programme, la constitution de corpus multimodaux (parole + vidéo) est également envisageable. Ce type de ressource est en effet très intéressant pour ce qui concerne les questions de référence.

2.2.2. Présentation technique : contenu des corpus distribués

Il était impératif que les corpus réalisés dans le programme *Parole Publique* répondent à un souci de standardisation permettant leur réutilisabilité par les chercheurs de la communauté[50]. C'est pourquoi j'ai défini une méthodologie de transcription[51] et de codage qui sera reprise tout au long du projet *Parole Publique*. En l'absence de norme établie, cette méthodologie cherche à se cadrer sur les conventions les plus répandues au sein de la communauté francophone. En particulier :

- La transcription suit les conventions qui ont été définies pour le français parlé par le GARS[52] et sont désormais reprises par l'équipe DELIC. Ces conventions ont été légèrement enrichies par

---

[48] On ne discutera pas ici du genre ou registre (*register* en anglais) auquel pourrait correspondre ces ressources. Rappelons simplement que le langage oral ne forme pas un tout et qu'il peut au contraire se caractériser par un continuum de variabilités (Biber D., 1988, Variations across speech and writing. Cambridge Univ. Press, Cambridge, MA.). Le « genre » sur lequel porte *Parole Publique* pourrait être caractérisé par une spontanéité d'élocution qui est renforcée par la forte interactivité des dialogues. Celle-ci se manifeste en particulier par une fréquence élevée de chevauchements. Le caractère finalisé de l'interaction se traduit enfin par une faible couverture lexicale.

[49] Hirschman L. (1998) Language understanding evaluations : lessons learned from MUC and ATIS, Actes *1st Conference on Language Resources and Evaluation*, *LREC'98*. Grenade, Espagne, 117-122. (p. 121-122).

[50] Je ne reviendrai pas sur l'importance de cette question et sur les efforts de standardisation ou de normalisation (TEI, standardisation ISO, etc.) menés par la communauté scientifique au cours de la dernière décennie : Romary L. (2000) Outils d'accès à des ressources linguistiques. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 193-212 ; Ide N. et Véronis J. (Dir.) (1999) Selected papers from TEI10 : celebrating the 10th anniversary of the Text Encoding Initiative, *Computers and the Humanities*, 33(1-2), Kluwer Academic Publ. , Dordrecht, Pays-Bas.

[51] Pour un aperçu de la variété des pratiques en matière de conventions de transcription et de ses incidences : Bilger M. (2000) Petite typologie des conventions de transcription de l'oral : quelques aspects pratiques et théoriques. *Cahiers de l'Université de Perpignan*, n° 31, Presses Universitaires de Perpignan, Perpignan, France. 77-92.

[52] Blanche-Benveniste C., Jeanjean C. (1987) Le français parlé : transcription et édition. Paris, Didier Erudition.

certaines recommandations EAGLES issues du projet SPEECHDAT[53]. Elle restent cependant en accord avec le principe d'objectivité défendu par Claire Blanche-Benveniste et ses collègues. En particulier, la transcription ne doit ainsi jamais être corrompue par des sur-interprétations de prononciation (*y'a ka*, *ya*, *chais pas*) ou par l'indication de faits para-linguistiques. Le détail des conventions de transcriptions est disponible sur le site Internet du programme.

- Grâce au financement de l'appel d'offre TECHNOLANGUE (cf. 2.2.3), les corpus distribués dans le cadre du programme *Parole Publique* seront enrichis à partir d'octobre 2003 par une annotation morphosyntaxique en parties du discours[54]. Cette annotation reprendra le jeu d'étiquettes qui a été défini au cours de l'action GRACE d'évaluation des étiqueteurs du français[55]. L'annotation sera réalisée de manière semi-automatique à l'aide du logiciel CORDIAL Analyseur de la société Synapse Technologies, ou d'une adaptation du système SIBYLLE réalisé par notre équipe (cf. chapitre 4).
- Chaque transaction est associé à un ensemble de descripteurs qui décrivent par exemple le contexte d'enregistrement, le thème de l'interaction, les caractéristiques des locuteurs etc. Dans le cadre d'un groupe de travail commun aux RTP 14 et 38 du CNRS (département STIC), je réfléchis actuellement avec d'autres collègues à la définition d'une base de descripteurs qui pourraient être utilisés à la fois en ingénierie des langues et en sciences du langage.

Les fichiers de transcription annotés sont encodés dans le format structuré XML (figure 2.3). Nous reprenons pour cela la DTD[56] définie par le logiciel libre *Transcriber*[57] que nous utilisons pour réaliser la transcription. Ce choix est motivé par le fait que *Transcriber* s'est imposé comme un standard de fait à côté des logiciels *Praat* et *WinPitch*.

```
< ?xml version= ”1.0 ” encoding= ”UTF-8 ” ?> < !DOCTYPE Trans SYSTEM “ trans-13.dtd ”>
<Trans scribe= ”Nicolas ” audio_filename= ”1ag0365 ” version= ”1 ” version_date= ”011008 ”>
<Speakers>
<Speaker id= ”spk1 ” name= ”hôtesse ” check= ”no ” type= ”female ” dialect= ”native ” accent= ” ”
scope= ”local ”/>
<Speaker id= ”spk2 ” name= ”client ” check= ”no ” type= ”female ” dialect= ”native ” accent= ” ” scope= ”local ”/
>
</Speakers>
<Topics>
<Topic id= ”to1 ” desc= ”1ag0365 ”/>
</Topics>
<Episode>
<Section type= ”report ” startTime= ”0 ” endTime= ”5.980 ” topic= ”to1 ”>
<Turn startTime= ”0 ” endTime= ”0.629 ” speaker= ”spk1 ”>
<Sync time= ”0 ”/>
bonjour madame
</Turn>
```

[53] Gibbon D., Moore R., Winski R. (Eds.) (1997) Handbook of standards and resources for spoken language systems, Berlin, Mouton de Gruyter, Berlin, Allemagne (recommandations définies en pp. 825-834).

[54] Cette annotation représente une « valeur ajoutée » essentielle à la mise en œuvre de traitements linguistiques dépassant la simple observation de cooccurrence de mots (Leech G., 1997, Introduction. *In* Garside R., Leech G., McEnery A., *Corpus annotation : linguistic information from computer text corpora*. Longman, London, UK. 1:18 ; Habert B., Nazarenko A., Salem A.,1997, Les linguistiques de corpus. Armand Colin, Paris, France).
D'une manière générale, l'utilisation de corpus annotés a montré son intérêt pour l'extraction de ressources terminologiques, le développement d'outils multilingues ou l'apprentissage de modèles de langage stochastiques (Véronis J., 2000, Annotation automatique de corpus : panorama et état de la technique. In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection I^{2}C. Hermès, Paris, France. 235 :250).

[55] Adda G., Mariani J., Paroubek P., Rajman M. et Lecomte J. (1999) L'action GRACE d'évaluation de l'assignation des parties du discours pour le français, *Langues*, 2(2), 119-129.

[56] DTD = *Document Type Definition*. Il s'agit d'une notion définie avec le langage de balisage structuré SGML. La DTD attachée à un document SGML décrit formellement l'organisation et la sémantique de l'information au sein de ce dernier. Cette notion se retrouve dans le langage XML. Pour une présentation rapide de SGML, XML et de leur utilisation en ingénierie des langues : Bonhomme P. (2000) Codage et normalisation de ressources textuelles. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 173-192.

[57] Barras C. *et al*. (1998) Transcriber : a free tool for segmenting, labeling and transcribing speech, Actes *1st Conference on Language Ressources and Evaluation*, *LREC'98*. Grenade, Espagne,, pp. 1373-1376.

```
<Turn speaker= ”spk2 ” startTime= ”0.629 ” endTime= ”3.420 ”>
<Sync time= ”0.629 ”/>
bonjour est ce que vous avez le programme de oui e e je
</Turn>
<Turn speaker= ”spk1 spk2 ” startTime= ”3.420 ” endTime= ”3.856 ”>
<Sync time= ”3.420 ”/>
<Who nb= ”1 ”/>
oui
<Who nb= ”2 ”/>
connaissances
</Turn>
<Turn speaker= ”spk2 ” startTime= ”3.856 ” endTime= ”4.24 ”>
<Sync time= ”3.856 ”/>
du monde
</Turn>
```

Figure 2.3 *— Exemple de transcription sans annotation morphosyntaxique ni descripteurs au format XML – DTD Transcriber (extrait du corpus OTG).*

La dernière version de la TEI permet désormais une représentation satisfaisante des corpus oraux[58]. Nous devrions donc nous tourner à court terme vers ce format standard de codage qui nous permettra d'intégrer de manière générique les descripteurs de corpus (entête TEI) et l'annotation morphosyntaxique des corpus.

Le codage XML est également traduit, pour des usages spécifiques, en format texte (ASCII). Ce codage conserve une structuration en tours de parole (figure 2.4) et peut intégrer une annotation morphosyntaxique en parties du discours. On remarquera que les chevauchements sont toujours représentés dans ce format.

```
                                fichier audio : 1ag0365
<001> hôtesse
    h: bonjour madame
<002> client
    c: bonjour est ce que vous avez le programme de oui e e je
<003> hôtesse+client
    h: oui
    c: connaissances
<004> client
    c: du monde
```

Figure 2.4 *— Exemple de transcription sans annotation morphosyntaxique ni descripteurs en format texte (extrait du corpus OTG identique à celui de la figure 2.3).*

Les corpus réalisés peuvent être librement récupérés sur le site du programme *Parole Publique*, sous réserve d'acceptation d'une convention d'utilisation peu contraignante. Ils sont également distribués dans le cadre de l'action spécifique ASILA et du projet ANANAS[59] du CNRS . Le paragraphe qui suit présente brièvement les corpus qui sont déjà disponibles.

2.2.3. Premiers résultats et action TECHNOLANGUE

Comme je l'avais précisé en introduction de ce mémoire (cf. chapitre 0, § 5.3), deux corpus de dialogue oral (*OTG* et *Ecole Massy*[60] : tableaux 2.1 et 2.2) sont d'ores et déjà diffusés dans le cadre

[58] Conventions de transcriptions de la parole de la TEI : http://www.tei-c.org/P4X/TS.html

[59] Le projet ANANAS (Annotation Anaphorique pour l'Analyse Sémantique de Corpus) a pour objectif de créer une base de corpus du français annotés en relations anaphoriques. Il est coordonné par Susanne Salmon-Alt (ATILF, Nancy). Site Toile : http://www.inalf.fr/ananas/site/htm/.

[60] Antoine J.-Y., Letellier-Zarshenas S., Nicolas P. , Schadle I., Caelen J. (2002) Corpus OTG et ECOLE_MASSY :
vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement. Actes *TALN'200*2, Nancy, France. vol. 1, 319-324.

du programme *Parole Publique*. Ils relèvent du domaine du renseignement touristique sur lequel portent nos travaux en compréhension de parole (cf. chapitre 4 § 1).

Tableau 2.1 *— Description synthétique des corpus OTG et Ecole Massy*

CORPUS	OTG	ECOLE MASSY
Durée d'enregistrement	117 minutes	45 minutes
Nombre de dialogues	315	31
Nombre de locuteurs	5 réceptionnistes / 315 touristes	1 enseignant / 19 élèves
Nombre de mots	25 695	5 300

Tableau 2.2 *— Distribution des dialogues des corpus OTG et Ecole Massy suivant leur durée*

Durée	< 30s	30s – 1 mn	1 mn – 2 mn	2 mn-3 mn	> 3 mn
OTG	294	77	36	2	0
Ecole Massy	2	6	16	7	0

Le **corpus *OTG*** (*Office du Tourisme de Grenoble*) a été constitué dans le cadre de l'ARC « Dialogue Oral » de l'AUF. Il a été enregistré par le laboratoire CLIPS-IMAG à la Maison du Tourisme de Grenoble. Les clients et l'agent n'ont été soumis à aucune consigne. La prise de son s'est effectuée en conditions réelles suivant une procédure semi-clandestine[61]. Il s'agit d'un corpus pilote que l'on peut estimer représentatif des usages réels dans le domaine du renseignement touristique. La transcription a été réalisée par notre équipe avec le soutien financier de l'AUF. Au total, 315 dialogues ont été transcrits, qui correspondent à 2 heures d'enregistrement. Ce corpus a une taille globale à 25 700 mots transcrits. A terme, il atteindra une taille critique de 40 000 mots et sera annoté en parties du discours.

Sensiblement plus petit (5300 mots pour 31 dialogues), le **corpus *Ecole Massy*** a été enregistré et transcrit sur fonds propres. Tout en relevant du renseignement touristique, les dialogues recueillis concernent une tâche plus précise de planification d'activités de loisirs. Ce corpus répond à une motivation scientifique spécifique : l'étude des usages langagiers chez de jeunes locuteurs. La population étudiée était constituée d'enfants de sept à huit ans enregistrés dans leur classe[62]. Etant donné ses motivations, ce corpus n'est pas utilisable pour la conception de systèmes standards de dialogue oral. L'adaptation des systèmes de dialogue oral à des publics spécifiques (personnes âgées[63], handicapés, adolescents, enfants…) représentera cependant une problématique importante dans les années à venir. Elle nécessitera alors le recours à des observations sur des corpus analogues à celui-ci. Ce corpus est également susceptible d'intéresser les chercheurs en sciences de l'éducation[64] et en psychologie du développement.

A court terme, les futures réalisations du programme *Parole Publique* se feront dans le cadre de l'appel d'offre *Technolangue* du ministère de la Recherche. Plus précisément, le groupe CORAIL intervient dans le projet AGILE-OURAL en qualité de fournisseurs de corpus oraux pour ce projet qui vise la distribution de composants de base pour le traitement automatique des langues (segmenteur, étiqueteur grammatical, outil de repérage d'entités, outil de segmentation thématique).

61 Ce n'est qu'à la fin de la transaction que le client était informé de l'enregistrement.

62 Le corpus *Ecole Massy* regroupe des dialogues simulés portant sur la tâche étudiée. Il a été enregistré dans une classe de CE1 d'une école primaire de Massy. Les consignes fournies aux enfants concernaient uniquement l'objectif de la transaction. Il s'agissait d'une recherche de séance de cinéma suivie de la planification d'une activité de loisirs sur la région parisienne. L'enseignant, qui jouait le rôle de l'agent, avait pour consigne de simuler un dialogue assez directif. Les transactions se sont faites sur les possibilités réelles de loisirs offertes au moment de l'enregistrement. A la demande de l'enseignant et à notre grand regret, les enregistrements ont été réalisés en notre absence.

63 Privat R. (2000), Interrogation multimodale de consultation de serveurs d'informations : application aux personnes âgées, Actes des *1ères Rencontres Jeunes Chercheurs en IHM*, *RJC-IHM'2000*, Ile de Berder, France, pp. 127-130.

64 Le Cunff C. (2002) De l'usage des corpus en didactique de l'oral : recherche et formation. Actes des *2ème journées de la Linguistique de Corpus*, Lorient, France. p. 25 (résumé).

Ces outils, qui sont développés en premier lieu pour le traitement de l'écrit, seront également validés sur les corpus de dialogue oral recueillis par les laboratoires VALORIA et SILEX. Ainsi, la transcription du corpus OTG a déjà été révisée et étendue dans le cadre du projet, et nous abordons maintenant la réalisation d'un corpus qui concernant l'accueil téléphonique (standard).

Au terme du projet, le laboratoire VALORIA disposera d'un corpus de dialogue oral qui comportera 200 000 mots transcrits et annotés et concernera les différents domaines d'application cités précédemment (§ 2.2.1). Au vu de la taille des corpus oraux déjà existants, cette ressource peut sembler de faible envergure. C'est oublier que la transcription de dialogue oraux très interactifs, comportant de nombreux chevauchements et interruptions, constitue une activité lourde et coûteuse. Si d'autres initiatives ne voient pas le jour (plate-forme corpus des RTP 14 et 38, par exemple), cette ressource devrait ainsi représenter, à la fin de l'année 2004, le plus grand corpus francophone de dialogue oral distribué librement.

Par sa diversité, ce corpus devrait nous permettre d'approfondir les analyses d'usages langagiers que j'ai menées jusqu'à présent avec le groupe CORAIL. Ce sont ces études que je vais maintenant présenter, en montrant leur intérêt pour la conception des systèmes de dialogue oral et plus généralement pour les recherches en ingénierie des langues.

3. LINGUISTIQUE DE CORPUS ET DIALOGUE HOMME - MACHINE

Bien que moins étudié que l'écrit, le langage oral a fait l'objet de multiples recherches linguistiques. Ces travaux ont cependant privilégié un langage parlé « général » d'où pouvait même être absente toute dimension interactive (monologues). A l'opposé, nos recherches en linguistique de corpus s'intègrent spécifiquement dans la perspective du dialogue homme - machine. C'est pourquoi nous nous concentrons sur l'étude de corpus pilotes de dialogue oral finalisé.

Ces corpus nous servent tout d'abord à prototyper les systèmes que nous réalisons. Le groupe CORAIL n'a pas encore les moyens de développer intégralement le cycle en « b » de conception que j'ai décrit précédemment (cf. § 1.4). Afin d'éviter les limitations d'une conception purement incrémentale, nous avons recours à deux types de ressources pour concevoir nos systèmes. Tout d'abord, nous utilisons des corpus d'amorçage qui sont analogues à ceux employés dans les approches par *bootstrap*. Il s'agit généralement de corpus simulés voire inventés. Les systèmes LOGUS et ROMUS présentés au chapitre 4 ont ainsi été prototypés à l'aide du corpus PARISITI réalisé par le LIMSI dans le cadre de l'ARC « Dialogue oral » de l'AUF.

Ces corpus artificiels sont toutefois mis en regard d'analyses préalables des usages qui portent cette fois sur des corpus pilotes. Dans la pratique, cette confrontation entre données à visées très pragmatiques et corpus d'études est peu formalisée. A l'usage, elle apparaît cependant féconde en terme de directions de conception. Il n'en reste pas moins que les enseignements que l'ont peut tirer de ces études restent fortement liées à une application particulière.

Nous retrouvons ici une limitation bien connue du dialogue oral homme - machine, à savoir son manque de généricité en terme de champs applicatifs couverts. Cette situation s'explique par l'effort important que nécessite le développement de tout nouveau système interactif. A défaut de pouvoir multiplier les systèmes opérationnels, l'étude différentielle de corpus pilotes portant sur des tâches variées peut fournir des renseignements utiles à la poursuite des recherches en dialogue oral.

C'est dans cette perspective que sont menées les analyses d'usage que je vais présenter dans les paragraphes suivants. Plus précisément, chaque étude tente de répondre, à partir d'une analyse de corpus pilotes, à une question d'ordre technologique. Ainsi :

- la première analyse de corpus (§ 3.1.) cherche à appréhender l'influence de la richesse sémantique de la tâche sur la complexité structurelle des énoncés oraux. Cette problématique est importante, puisqu'elle concerne la généralisation du dialogue oral homme-machine à des domaines d'application plus complexes. Elle pose la question de la nécessité d'un recours à des techniques d'analyse linguistique plus fines en CHM orale.

- la seconde étude (§ 3.2.) porte sur l'analyse des variations d'ordonnancement linéaire dans des contextes interactifs variés. Centrée sur l'étude des dislocations en français parlé, son objectif est de répondre à une interrogation précise : est-il envisageable d'utiliser des analyseurs syntaxiques non projectifs[65] pour la compréhension de parole en situation de dialogue finalisé. Comme nous le verrons, la portée de cette étude concerne néanmoins le traitement automatique du français parlé dans sa globalité.
- la troisième étude (§ 3.3.) concerne l'analyse de réparations orales (répétitions, corrections, faux départs) sur des corpus pilotes correspondant à des tâches différentes. Son objectif est de vérifier si les techniques de détection par patrons de ces procédés spécifiques à l'oral spontané peuvent être utilisables dans tous les contextes d'application.
- la dernière étude (§ 3.4.), qui porte sur la co-référence anaphorique à l'oral, suit actuellement son cours. Il s'agit également d'une étude prospective qui cherche à vérifier si les méthodes superficielles de résolution des anaphores qui ont été développées pour l'écrit peuvent s'appliquer à l'oral spontané.

Les résultats que je vais présenter dans les paragraphes suivants montreront en quoi ces analyses d'usages peuvent être utiles à la conduite des recherches en ingénierie des langues. On notera que ces études amonts sont également susceptibles d'intéresser la linguistique de corpus.

3.1. Complexité de la tâche et complexité du langage oral

Contrairement aux études de corpus qui seront présentées dans les paragraphes suivants, les travaux qui nous intéressent ici n'ont pas fait l'objet d'une analyse quantitative. Les conclusions qualitatives que je vais présenter ne pourront donc pas être vérifiées sur d'autres corpus. Il me semble cependant important de présenter ces travaux qui ont guidé la conception du système de compréhension ROMUS (cf. chapitre 4 § 1) et qui ont été à l'origine de mon implication en linguistique de corpus.

Comme je l'expliquerai au cours du chapitre 4, les systèmes de dialogue actuels tirent parti du caractère finalisé de l'interaction pour conduire des analyses relativement robustes. Dans la perspective d'une généralisation de la communication orale homme-machine, je me suis interrogé très tôt sur les possibilités d'adaptation de cette démarche *ad hoc* à des tâches plus complexes. Un des objectifs du doctorat de Jérôme Goulian était précisément de défricher cette problématique au niveau de la compréhension automatique de la parole.

Le champ d'application retenu pour ce doctorat concernait le renseignement touristique. Mon premier souci a été de situer la complexité de ce domaine par rapport à celle d'autres applications déjà étudiées. Il importait de distinguer la complexité du domaine d'application proprement dit de celle de la tâche mise en jeu. J'ai choisi de caractériser la richesse du domaine par le nombre de concepts sémantiques nécessaire à la représentation de l'univers de la tâche[66]. Dans le cadre du dialogue homme - machine, un concept est une classe sémantique qui regroupe tous les lexèmes exprimant une même unité de sens du point de vue de l'application. Par exemple, le renseignement aérien implique des concepts tels que le [numéro du vol], la [date de départ] ou l'aéroport de [correspondance].

L'analyse de la littérature a permis de relever quelques exemples de complexité évalués par différents auteurs (tableau 2 .3 page suivante). Ces mesures ont été obtenues de manière indépendante. Intuitivement, elles semblent toutefois donner des indications cohérentes. Par exemple, la recherche d'horaire de trains (ARISE-IRIT) apparaît sensiblement moins complexe que

[65] Cette notion sera présentée ultérieurement (cf. § 3.2).

[66] Cette définition me semble plus pertinente que celle reposant sur la taille du lexique. En effet, une application peut mobiliser un vocabulaire conséquent tout en restant assez simple. C'est le cas du renseignement ferroviaire, où les nombreux noms de villes correspondent à un nombre très restreint de concepts (gare de départ ou d'arrivée). La notion de perplexité définie par le TAL probabiliste est de peu d'utilité ici, puisqu'elle évalue conjointement la complexité d'une tâche (ou d'un domaine) et l'efficacité du modèle de langage utilisé pour la décrire. Il aurait été intéressant d'appréhender plus finement encore la complexité de la tâche en considérant, par exemple, les relations entre concepts sémantiques. Ces informations ne sont malheureusement jamais présentées dans la littérature.

le domaine général du renseignement ferroviaire (MASK ou ARISE-LIMSI). On retrouve également une proximité attendue entre renseignements aérien et ferroviaire.

Comme le montre le tableau 2.3, le renseignement touristique présente une richesse sémantique qui est significativement plus importante que celle d'autres domaines d'application usuels (ATIS, ARISE, etc.). La question qui se pose alors est de savoir si cette complexité se retrouve au niveau des productions orales des interlocuteurs. En particulier, nous nous sommes demandés si les énoncés rencontrés dans le domaine du renseignement touristique ne présentaient pas une structure plus complexe qui aurait nécessité le recours à des traitements linguistiques fins.

La complexité structurelle d'un énoncé peut être caractérisée par la profondeur de son arbre de dépendances[67]. Cette mesure objective nécessite cependant la constitution d'une banque d'arbres (*treebanks* en anglais) sur des corpus représentatifs. Le corpus arboré réalisé par le laboratoire TALANA, qui n'était pas disponible au moment de l'étude, ne concerne de toute façon pas le dialogue oral.

Tableau 2.3 *— Complexité d'une tâche en dialogue oral finalisé évaluée en terme de richesse sémantique (nombre de concepts + nombre de schémas) : quelques exemples*

Domaine d'application	Auteur	Laboratoire	Complexité
Horaires de train (ARISE)	Bousquet[68]	IRIT	29 — 32
Renseignement aérien (ATIS)	Minker[69]	LIMSI	43
Renseignement ferroviaire (MASK)	Minker[70]	LIMSI	52
Renseignement ferroviaire (ARISE)	Maynard, Lefèvre[71]	LIMSI	72
Renseignement touristique	Goulian, Antoine[72]	VALORIA	151
Renseignement touristique et planification de voyage (NESPOLE)	Rossata, Blanchon, Besacier[73]	CLIPS	276

Nous nous en sommes donc remis à une étude qualitative sur des corpus de dialogue oral correspondant à des domaines plus ou moins riches. Cette analyse différentielle ne nous a pas permis d'observer de différences significatives de complexité structurelle. En particulier, le corpus Levelt[74], qui relève d'un domaine d'application très restreint (conception de dessins géométriques),

[67] D'autres métriques peuvent être considérées, comme la longueur moyenne des énoncés ou la fréquence d'utilisation des propositions subordonnées (Biber D., 1986, Spoken and written textual dimensions in English : resolving the contradictory findings. *Language*, 62(2), 384-414). Ces métriques me semblent néanmoins trop ponctuelles.

[68] Il s'agit d'une application définie dans le cadre du projet européen ARISE. Deux modèles de langage ont été étudiés (ML1 et ML2), qui reposent sur des concepts de granularités différentes : Bousquet-Vernhettes C. (2002) Compréhension robuste de la parole spontanée dans le dialogue oral homme - machine : décodage conceptuel stochastique. Thèse de l'Université Paul Sabatier, Toulouse, France, 26 septembre 2002. (Données page 81).

[69] Minker W. (1995) An English version of the LIMSI L'ATIS System. Rapport LIMSI 95-12. Orsay, France.

[70] Minker W. (1999) Compréhension automatique de la parole. L'harmattan, Paris, France

[71] Par opposition au système ARISE de l'IRIT, la tâche étudiée ici est le renseignement ferroviaire en général : Maynard H., Lefèvre F. (2002) Apprentissage d'un module stochastique de compréhension de parole. Actes *XXIV° Journées d'Etudes sur la Parole*, JEP'2002, Nancy, France. 129-132.

[72] Goulian J. (2000) Analyse linguistique détaillée pour la compréhension automatique de la parole spontanée, Actes *RECITAL'2000*, Lausanne, Suisse.

[73] Le projet Nespole ! concerne la traduction de parole (communication médiée par ordinateur) sur une tâche de renseignement touristique et de planification de voyage : Rossato S., Blanchon H., Besacier L. (2002) Evaluation du premier démonstrateur de traduction de parole dans le cadre du projet Nespole ! Actes *TALN'2002*, *Conférence associée " couplage de l'écrit avec l'oral*, Nancy, France. vol 2., 149-161.

[74] Le corpus Levelt a été recueilli à l'Institut de la Communication Parlée. Il correspond à une application de conception de dessins assistée par ordinateur. Le système, simulé par magicien d'Oz, se contente de réagir aux directives du locuteur. La partie du corpus qui a été considérée ici regroupe deux tâches destinées à étudier les procédés de description de configurations spatiales. Dans une première tâche, le locuteur doit décrire les actions à mener pour reproduire à l'écran des figures géométriques abstraites (figures de Levelt). La seconde tâche consiste à décrire des figures représentant des plans d'intérieur dont l'organisation géométrique reprend celles de la première tâche : Ozkan H., Bisseret A., Caelen J. (1991) Analyse de dialogues finalisés dans une perspective communicationnelle Actes des *3èmes journées sur l'ingénierie des Interface Homme - machine*, *IHM'91*, Dourdan,

comporte de nombreux énoncés complexes. En témoigne l'énoncé (2.1) où l'on remarque l'usage successif de deux subordonnées relatives :

(2.1) *alors on reprend un carré qu'on met à gauche et en bas de l'écran un trait vertical qui part du centre de la face supérieure de ce carré* (Levelt.6)

Ce résultat recoupe l'affirmation de Gérard Sabah et ses collègues selon laquelle la finalisation du dialogue ne réduit pas la complexité des phénomènes linguistiques rencontrés[75]. Ce que Jean-Marie Pierrel et Laurent Romary résument ainsi[76] :

« limiter le domaine d'application ne réduit pas les phénomènes linguistiques à traiter, que ce soit pour les caractéristiques de la langue ou les traitements de l'implicite inhérent à tout dialogue »

Cette confirmation expérimentale masque une autre influence de la complexité du domaine d'application. Lorsqu'on quitte le dialogue fortement finalisé, l'accroissement du nombre de concepts sémantiques que nous avons observé dans le tableau 2.3 se traduit par l'apparition d'ambiguïtés sémantiques[77]. Je reviendrai ultérieurement (cf. chapitre 4 § 1.3) sur l'influence de cette ambiguïté en compréhension automatique de la parole. Pour l'heure, on notera que c'est l'observation sur corpus de multiples cas d'ambiguïté qui nous a conduit à intégrer *in fine* des traitements linguistiques fins dans le système de compréhension ROMUS.

Les trois études de corpus que je vais maintenant présenter ont également une influence sur la conception de nos systèmes. Par leur nature quantitative, elles témoignent encore plus clairement de l'intérêt de ces études prospectives.

3.2. Variabilité linéaire et CHM orale : influence de l'interactivité du dialogue

Cette seconde étude s'intègre également dans le cadre de la conception du système ROMUS. Les travaux précédents nous ayant renseigné sur l'intérêt d'une modélisation linguistique fine, j'ai proposé d'adapter le formalisme des grammaires de liens pour le traitement du langage parlé (cf. chapitre 4, § 1.5.). Une des particularités de ce formalisme est de ne pas pouvoir traiter les dépendances discontinues. Il était donc intéressant d'évaluer en amont de la conception les conséquences réelles de cette limitation.

3.2.1. Variabilité linéaire et ingénierie des langues

La variabilité de l'ordre des mots dans l'énoncé et son corollaire, l'apparition de dépendances discontinues, a de tous temps interpellé la théorie linguistique comme le traitement automatique des langues. A la suite des travaux fondateurs de Lucien Tesnière, cette question a constitué un des appuis théoriques majeurs des grammaires de dépendances face aux théories chomskyennes[78]. Le traitement des structures discontinues fut également une des motivations de la création des grammaires syntagmatiques liées par la tête[79]. Il a enfin été étudié[80] dans le cadre des grammaires d'arbres adjoints (TAG).

France, 175-186).

75 Sabah G., Vivier J., Vilnat A., Pierrel J-M., Romary L., Nicolle A. (1997) Machine, langue et dialogue. L'Harmattan, Paris, France.

76 Pierrel J-M., Romary L. (2000) Dialogue homme - machine. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 331-350. Citation page 333-334.

77 Cet accroissement d'ambiguïté se retrouve partiellement dans l'augmentation de perplexité que l'on relève entre le renseignement aérien (perplexité de 20 pour un trigram) et l'application *Wall Street Journal* (informations financières : perplexité de 128 pour un trigram) : Huang X., Acero A., Hon H.-W. (2001) *op. cit*. (pages 561).

78 Tesnière L. (1959) Eléments de syntaxe structurale, Klincksiek, Paris, France ; Covington M. (1990) Parsing discontinuous constituents in dependency grammar. *Computational Linguistics*, 16(4), 234-236 ; Hudson R. (2000) Discontinuity. *Traitement Automatique des Langues*, *TAL*, 41(1), 15-56, Hermès, Paris.

79 Pollard C., Sag I.(1994) Head-driven Phrase Structure Grammar. University of Chicago Press, Chicago, Michigan.

80 Rambow O., Joshi A. (1994) A formal look at dependency grammars and phrase-structure grammars with special considerations of word-order phenomena, *In* Wanner L. (ed.), *Current issues in Meaning-Text Theory*, Pinter, Londres, Royaume-Uni.

Dans la perspective du traitement automatique de la langue, on peut distinguer deux niveaux de variabilité de l'ordre linéaire[81] :

- la **variabilité faible** correspond à des déplacements de constituants qui n'induisent pas de discontinuité dans la structure de l'énoncé. C'est le cas du mouvement du groupe prépositionnel *pour Rio Sao Paulo* dans l'énoncé (2.2) suivant :

  (2.2) *bon sinon **pour Rio Sao Paulo** je pense qu'il y a pas mal de vols* (Air France.II.8.C68)

- la **variabilité forte** se traduit au contraire par la production d'énoncés discontinus, que l'on qualifie également de **non-projectifs**. Dans l'énoncé (2 .3) ci-dessous, le déplacement de l'adverbe *maintenant* casse la dépendance entre la subordonnée relative *qui est nouveau* et son antécédent *un tarif encore plus intéressant sur Londres* :

  (2.3) *vous savez on a un tarif encore plus intéressant sur Londres **maintenant** qui est nouveau* (Air France.II.33.O17)

La distinction entre variabilité forte et faible interroge l'ingénierie des langues dans ses fondements. D'une part, certains formalismes syntaxiques tels que les grammaires de liens reposent sur le postulat de la projectivité du langage[82]. D'autre part, si la variabilité faible peut être traitée dans un cadre hors contexte, sa modélisation induit une augmentation pénalisante d'ambiguïté structurelle[83]. Notons enfin, comme le fait remarquer Covington[84], que la question de la variabilité linéaire est d'autant plus pressante que l'on a affaire à des langues à ordre variable (russe, finnois, tchèque, etc.). Un des objectifs de l'étude présentée ci-dessous est précisément de vérifier si le français parlé en situation de dialogue finalisé est un langage à ordre aussi rigide que le français écrit « standard ».

3.2.2. Etude de corpus

Cette étude a été menée sur deux corpus relevant de champs d'application différents (tableau 2.4) :

- Le corpus **Murol** a été recueilli par le laboratoire CLIPS-IMAG[85]. Il réunit des conversations téléphoniques simulées entre deux compères qui jouent respectivement le rôle du touriste et de l'agent d'un syndicat d'initiative. Les scénarii sur lesquels porte l'interaction concernent à la fois des problèmes de localisation (mise à jour de plan) et la planification de diverses activités de loisir (renseignement touristique). L'univers de la tâche est donc relativement riche.
- Le corpus **Air France** a été recueilli par l'équipe de Marie-Annick Morel (Université de la Sorbonne Nouvelle). Il a ensuite été retravaillé par Pierre Nerzic dans le cadre du projet DALI[86]. Il réunit des conversations téléphoniques réelles entre le centre de réservation de la compagnie Air France et des particuliers ou des agences de voyage. Tout en s'inscrivant dans le cadre du renseignement aérien, la tâche concernée est plus riche que celle étudiée dans le programme ATIS. Ce domaine d'application présente donc une complexité relativement modérée.

Afin de discriminer les sources de variabilités, j'ai considéré à la fois le degré d'interactivité des dialogues et la complexité de la tâche. Cette dernière dimension m'a conduit à scinder en deux le corpus Murol en fonction de la sous-tâche imposée par le scénario (tableau 2.4). La sous-tâche de mise à jour de plan est très restreinte, puisqu'elle se résume à la modification de certaines données

---

81 Holan T., Kubon, Oliva K., Plátek M. (2000) On complexity of word order, *Traitement Automatique des Langues*, *TAL*., 41(1), 273-300, Hermès, Paris.

82 Sleator D. D. K., Temperley D. (1991) Parsing English with a link grammar , *rapport de recherche CMU-CS-91-196*, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA.

83 Il s'agit bien entendu d'une fausse ambiguïté (*spurious ambiguity*) due à l'analyse. De ce point de vue, on contestera la position de Holan et ses collègues (*op. cit*.) qui relativisent les conséquences de la variabilité faible sur le traitement automatique du langage.

84 *Op. cit* ; on consultera également l'article de Holan T. *et al.* (2001) (op. cit.) sur la complexité des structures non-projectives en anglais, comparée à celle du tchèque.

85 Bessac M., Caelen G. (1995), Analyses pragmatiques, prosodiques et lexicales d'un corpus de dialogue oral homme - machine, Actes *JADT'95*, Rome, Italie, 363:370.

86 Sabah G. (1994) Projet DALI, *rapport d'activité GDR-PRC CHM*, 71-88. Disponible sur la Toile à l'adresse suivante : http://www-geod.imag.fr/pages_html/projets/DALI.htm

toponymiques. Une fois ce plan actualisé, les interlocuteurs devaient aborder une seconde sous-tâche relevant du renseignement touristique en général (complexité relativement importante).

Tableau 2.4 *— Caractéristiques des corpus d'étude.*

Corpus	Tâche	Type de dialogue	Interactivité	Complexité	Taille
Murol	mise à jour plan	H-H simulé	très forte	faible	1078 tours
	renseign^{mt} touristique	H-H simulé	très forte	significative	de parole
Air France	renseign^{mt} aérien	H-H réel (pilote)	assez forte	modérée	5179 tours

Le degré d'interactivité a été évalué de manière subjective par observation de la fréquence des chevauchements et des interruptions. Si l'interactivité est réelle pour le corpus Murol comme pour le corpus Air France, le dialogue reste cependant plus contenu dans le second cas. Le degré d'interactivité a donc été jugé moins important dans ce cas.

Le recensement des extractions a été effectué manuellement[87]. Chaque observation a été caractérisée suivant une typologie précise :

- *direction de l'extraction* — antéposition ou postposition
- *procédé mis en jeu* — inversion, dislocation, présentatifs, énoncés binaires[88].
- *fonction syntaxique de l'élément extrait* — sujet, argument, modifieur ou associé[89].
- *projectivité de l'extraction* — extraction discontinue (variabilité forte) ou non (faible).

Nous avons finalement mesuré la fréquence d'apparition et la distribution de ces différents types d'extractions dans chaque corpus. L'unité de segmentation choisie pour ces calculs de fréquence est le tour de parole, défini comme toute partie du dialogue au cours duquel l'identité et le nombre des locuteurs (chevauchements) restent constants.

3.2.3. Résultats et enseignements : analyse non-projective du langage parlé

On peut tirer plusieurs enseignements de cette étude[90]. Tout d'abord, on notera que si le français est une langue à ordre fixe, le problème que posent les extractions au traitement automatique ne saurait être ignoré dans le cas du dialogue oral finalisé. Le tableau 2.5 (page suivante), qui présente la fréquence d'apparition moyenne des extractions sur nos corpus, montre au contraire que ces procédés sont très répandus à l'oral. Par exemple, environ un quart des énoncés du corpus Murol comporte au moins un élément détaché.

On observe par ailleurs des différences significatives entre les trois corpus. Une analyse statistique montre que cette variation est expliquée par le degré d'interactivité du dialogue et non par la complexité de la tâche.

87 L'extraction automatique de procédés complexes constitue une problématique de recherche à part entière qui a donné peu de résultats jusqu'ici. Voir par exemple le travail de Sandrine Caddeo et Christophe Benzitoun sur l'apposition : Benzitoun C., Caddeo S. (2002) La recherche automatique des appositions, Actes *2èmes journées de la linguistique de corpus*, Lorient, France, p. 8 (résumé).

88 Cette classification reprend celle utilisée dans : Gadet F. (1992) Le français populaire, PUF, Paris, France.

89 Cette classification reprend celle utilisée par Claire Blanche-Benveniste dans ses études grammaticales : Blanche-Benveniste C. (1997) Approches de la langue parlée en français, Coll. *L'essentiel Français*, Ophrys, Paris, France.

90 Pour une présentation exhaustive des résultats avec validation statistique, on consultera : Antoine J-Y., Goulian J. (2001) Etude des phénomènes d'extraction en français parlé sur deux corpus de dialogue oral finalisé. Application à la communication orale homme - machine. *Traitement Automatique des Langues*, *TAL*, 42(2), 413-440

***Tableau 2.5** — Fréquence d'apparition des extractions sur chaque corpus (nombre moyen de tours de parole présentant au moins une extraction).*

Corpus	Tâche	Complexité	Interactivité	Fréquence moyenne	Ecart-type
Murol	mise à jour plan	faible	très forte	20,5 %	11,1 %
	renseign^{mt} touristique	importante	très forte	28,5 %	10,1 %
Air France	renseign^{mt} aérien	modérée	Assez forte	13,6 %	10,5 %

Cette conclusion est relativement intuitive. Une interactivité élevée traduisant un fort engagement des interlocuteurs, il est compréhensible que ces procédés de mise en relief soient plus utilisés. Un dialogue homme - machine plus coopératif se traduisant nécessairement par une interactivité accrue, il est à prévoir que le problème des extractions se posera avec encore plus d'acuité à l'avenir.

On notera ensuite qu'une remarquable régularité structurelle préside à la réalisation des extractions. Quelle que soit la caractéristique étudiée — sens du détachement (tableau 2.6), nature du procédé utilisé (tableau 2.7), nature de l'élément déplacé (tableau 2.8) — les études distributionnelles qui ont été réalisées ne présentent pas de différences statistiquement significatives entre les corpus.

***Tableau 2.6** — Distribution des extractions en fonction du sens du détachement*

Corpus	antépositions	postpositions	*écart-type*
Murol	82,5 %	17,5 %	σ = *20,4 %*
Air France	85,5 %	14,5 %	σ = *8,7 %*

***Tableau 2.7** — Distribution des extractions en fonction du procédé*

Corpus	inversion	double-marquage	présentatif	élément binaire
Murol	60,6 %	24,9 %	13,2 %	1,3 %
Air France	67,8 %	16,8 %	14,4 %	1,0 %

***Tableau 2.8** — Distribution des extractions suivant la fonction syntaxique de l'élément détaché*

Corpus	sujet	argument	modifieur	associé
Murol	30,7 %	12,0 %	27,4 %	30,0 %
Air France	25,4 %	5,3 %	23,5 %	45,8 %

Cette régularité s'explique certainement par l'existence de contraintes normatives qui limitent les possibilités de déplacement en français. Françoise Gadet relève ainsi l'influence de l'ordre canonique Sujet-Verbe-Objet sur la direction de l'extraction[91]. Cette observation se retrouve dans cette étude : dans tous les cas, plus de 90% des extractions respectent l'ordre SVO.

On peut considérer à la suite de Françoise Gadet[92] que « *la langue comme une variation réglée : elle fait système malgré l'hétérogénéité* ». Catherine Kerbrat-Orecchioni parle de même de liberté surveillée de l'oral dans l'interaction[93]. Du point de vue de l'ingénierie des langues, cette régularité est importante. Elle ouvre en effet la porte à un traitement générique des phénomènes d'extractions.

C'est toutefois un résultat non trivial qui a le plus retenu notre attention. Il apparaît en effet que les extractions non projectives sont rares, avec pour conséquence une proportion très marginale d'énoncés discontinus (tableau 2.10). Ainsi, les corpus étudiés ne présentent que de très rares exemples de relativisations à dépendance non bornée, sur lesquelles s'attardent pourtant nombre de chercheurs en TAL[94]. De même, rares sont les exemples de discontinuités dus à l'extraction de mots-question, procédés pourtant largement étudiés dans la littérature anglo-saxonne (*wh-question*).

91 Gadet F. (1992) *op. cit*.
92 Gadet F. (1989) Le français ordinaire. Armand Colin, Paris, France. (citation page 181).
93 Kerbrat-Orecchioni C. (1999) *op. cit*.

Tableau 2.10 *— Fréquence d'apparition des extractions non projectives (pourcentage par rapport à l'ensemble des extractions observées et par rapport à l'ensemble des énoncés)*

Corpus	% d'extraction non projectives	% d'énoncés discontinus
Murol	0,5 % (σ = 0,6 %)	0,2 % (σ = 0,2 %)
Air France	2,3 % (σ = 7,4 %)	0,4 % (σ = 0,9 %)

On peut se demander si nous sommes en présence d'une spécificité de la parole spontanée en situation de dialogue finalisé, ou si ce constat peut être étendu au français en général. Quoiqu'il en soit, les résultats de cette étude remettent en cause certaines préoccupations prétendument prioritaires du traitement automatique des langues.

Cette analyse des usages montre que le choix des modèles linguistiques utilisés en dialogue oral homme - machine ne dépend pas de la question de la variabilité forte du langage étudié. Elle démontre le caractère prédictif de ces études amonts en nous renseignant sur la pertinence de l'emploi d'un formalisme non-projectif (grammaires de liens) pour l'analyse du langage parlé.

3.3. Procédés de l'oral spontané et traitements linguistiques de surface

L'étude que je vais maintenant présenter a une portée moins étendue que la précédente. Elle concerne néanmoins une question essentielle pour la CHM orale, puisqu'elle s'intéresse à la caractérisation des réparations dans les énoncés oraux spontanés.

3.3.1. Réparations : description linguistique

Comme tous les procédés de l'oral spontané (hésitations, incises, interruptions, etc.), les réparations sont la manifestation d'une production en direct du message parlé. Claire Blanche-Benveniste les qualifie ainsi de « *traces de production* » de la parole[95]. Elles témoignent d'une recherche ou d'un ajustement de dénomination au cours de l'élocution. On distingue généralement trois types de réparations : répétitions, (auto)corrections et faux départs.

Les **répétitions** peuvent tout d'abord servir à meubler l'attente due à une recherche de dénomination. Le locuteur se répète sans chercher à les corriger, avant de reprendre le fil de son discours. Ces répétitions peuvent être multiples, comme dans l'exemple (2.4) ci-dessous. Dans la plupart des cas, la répétition ne se limite toutefois pas à une simple reprise. Elle se traduit au contraire par une reformulation qui vise à préciser la dénomination, comme dans le cas du procédé d'enrichissement lexical décrit par l'exemple (2.5).

(2.4) *vous avez le vous avez le le le montant du billet e le montant de e du billet sur la souche* (Air France.I.4.O16)

(2.5) est-ce que vous pourriez m'indiquer où je pourrais me trouver où je pourrais me procurer des des prospectus e des petits de la documentation écrite (Air France.II.14.C3)

(2.6) *il y a des sculptures sur bois qui sont inté- très intéressantes* (Murol.3)

Notons enfin que le point d'interruption de la répétition peut intervenir en milieu de mot. Elle est alors marquée par une amorce lexicale[96] comme dans l'exemple (2.6).

Dans tous les cas de figure, on remarque que les répétitions se manifestent par un travail sur l'axe paradigmatique. Le locuteur se reprend généralement au début du syntagme sur lequel s'effectue ce travail de dénomination. L'analyse en grille développée par Claire Blanche-Benveniste et ses collègues du GARS rend bien compte de cette construction (figure 2.5).

---

94 Kahane S. (2000) Extractions dans une grammaire de dépendance lexicalisée à bulles, *Traitement Automatique des Langues, TAL*, 41(1), Hermès, Paris, France. 211-244.

95 Blanche-Benveniste C., Bilger M., Rouget C., Van den Eynde K. (1990) Le français parlé : études grammaticales. CNRS, Paris, France (p. 17)

96 JeanJean C. (1984) Les ratés c'est fa- fabuleux, *Linx*, 10, 171-177

est-ce que vous pourriez m'indiquer où je pourrais me trouver
où je pourrais me procurer des
des prospectus e
des petits
de la documentation écrite

***Figure 2.5** — Analyse en grille de l'énoncé (2.5)*

Les **corrections** se manifestent également sur l'axe paradigmatique. A la différence des répétitions, elles ne traduisent pas une simple recherche de dénomination, mais la correction d'une production erronée. Elles peuvent aussi refléter l'ajustement d'une dénomination qui semble imprécise au locuteur. Levelt parle dans ce cas d'« *appropriateness repairs* » plutôt que de correction[97].

Dans tous les cas, il est difficile de distinguer répétitions et corrections sur des critères purement syntaxiques. On note parfois la présence d'un terme d'édition (cf. § 3.3.2) qui marque la correction. Même dans ce cas, la distinction reste assez subjective, comme le montre l'exemple (2.7) :

(2.7) *un carré à gauche non quand même pas tout à fait à gauche* (Levelt.5)

Les **faux départs** correspondent enfin à un type particulier de répétition ou de correction. Dans ce cas, la réparation ne se limite pas à un ou plusieurs syntagmes : c'est l'ensemble de l'énoncé déjà prononcé, groupe verbal compris, qui est repris. Le début de l'énoncé (2.4) peut ainsi être analysé comme un faux départ. Du fait de leur étendue, ces réparations sont les plus difficiles à détecter automatiquement.

3.3.2. Réparations et traitement du langage parlé

Comme l'ont montré les exemples précédents, les réparations peuvent présenter une organisation relativement complexe où les segments repris viennent enrichir la caractérisation du concept que le locuteur cherche à exprimer[98]. Ces constructions posent de sérieux problèmes à l'analyse automatique car elles perturbent la structure des énoncés. Elles gênent en particulier fortement la reconnaissance et la compréhension automatique de la parole. Il est donc important de mieux connaître leur fonctionnement afin d'élaborer des systèmes qui soient plus tolérants à leur présence.

Les réparations ont fait l'objet de nombreuses études linguistiques. A l'image des recherches du GARS, ces travaux se limitent toutefois à des études descriptives qu'il est difficile de traduire d'un point de vue informatique. C'est pourquoi les systèmes actuels préfèrent s'en remettre à des stratégies d'analyse partielle plutôt que d'attaquer de front le problème de la détection des réparations.

Certains chercheurs ont pourtant commencé à s'intéresser à cette question depuis une dizaine d'années. La plupart de ces recherches reposent sur une approche qui a été proposée par Bear, Dowding et Shriberg[99]. A la suite des travaux de Levelt[100], les réparations sont tout d'abord décrites comme la succession d'un *reparandum* (partie de l'énoncé qui sera corrigée par la suite), d'un *terme d'édition* optionnel (marqueur de reprise) et d'une altération (partie de l'énoncé qui corrige ou complète le reparandum). Par exemple:

(2.8) *le montant* [*de*]$_{\text{reparandum}}$ [*euh*]$_{\text{édition}}$ [*du billet*]$_{\text{altération}}$

On définit alors des patterns de détection superficiels qui reposent sur l'identification de suites de

97 Levelt W. (1983) Monitoring and self-repair in speech, *Cognition*, 14. 41-104.

98 Okada et Otsuka parlent ainsi d'élaboration incrémentale de la dénomination : Okada, H. Otsuka (1993) Incremental elaboration in generating spontaneous speech, actes *International Symposium on Spoken Dialogue*, *ISSD'93*, Tokyo, Japon, 49-52

99 Bear J., Dowding J., Shriberg E., Price P. (1993) A system for labeling self-repairs in speech. *SRI Technical note S22*.

100 Levelt W. (1983) *op. cit* ; Bruno Martinie adopte lui aussi une description linéaire, tout en étant conscient des limites (relevée par les chercheurs du GARS) de cette approche : Martinie B. (2001) Remarques sur la syntaxe des énoncés réparés en français parlé. *Recherches sur le Français Parlé*, 16 (2001), 189-206.

mots du reparandum et de l'altération qui se répètent de manière identique (M), sont repris (mots différents jouant le même rôle syntaxique : R) ou sont ajoutés (X). S'y ajoute éventuellement un terme d'édition noté ET. Reprenons par exemple l'exemple (2.4) :

(2.4) *vous avez le vous avez le le le montant du billet e le montant de e du billet sur la souche*

Cinq patterns de réparation peuvent être définis sur cet énoncé (la barre verticale marque le point d'interruption) :

1)	M1 M2 M3 | M1 M2 M3	*vous avez le* | *vous avez le*
2)	M1 | M1	*le* | *le*
3)	M1 | M1	*le* | *le*
4)	M1 M2 R3 X | ET M1 M2 R3	*le montant du billet* | *e le montant de*
5)	R1 | ET R1 M2	*de* | *e du billet*

L'observation de ces patterns nous renseigne sur les limites de ces approches superficielles :

- leur manque de précision les prédisposent aux risques de surgénérativité. Par exemple, le pattern M1 M1 peut être déclenché par erreur avec un verbe pronominal : [*vous vous*] *trompez* ,
- leur nature superficielle les limitent à des constructions assez simples. En particulier, ils ne modélisent que partiellement les réparations enchâssées (patrons 4 et 5).
- ils ne font pas de distinction entre répétition et correction.

Ces limitations expliquent le manque de fiabilité les systèmes actuels de détection par pattern-matching (tableau 2.11 page suivante). Ces résultats sont cependant comparables au comportement des rares systèmes de détection reposant sur une analyse syntaxique profonde[101].

En conclusion, il est clair que ces méthodes superficielles ne peuvent résoudre l'ensemble des difficultés que pose la parole spontanée. Je me suis cependant demandé si elles ne pourraient pas être utiles dans une perspective moins ambitieuse qui sacrifierait le rappel au profit d'une précision maximale. L'idée étant qu'une détection sans fausse alerte d'une partie des réparations peut réduire le coût computationnel des étapes ultérieures d'analyse.

A ma connaissance, aucune étude n'a été menée sur la détection des réparations par *pattern-matching* en français parlé[102]. Plutôt que de développer un système sur un domaine spécifique, il m'est apparu plus intéressant de procéder à une étude préalable de faisabilité sur différents corpus pilotes. Le propos de cette analyse de corpus était le suivant : étudier la distribution des patterns de réparations sur différents corpus de dialogue oral finalisé pour estimer l'adéquation de ces approches avec les besoins de la CHM orale.

Tableau 2.11 *— Comparaison des performances de certains systèmes de détection des réparations*

[101] Avec son analyseur à base de grammaires d'arbres adjoints lexicalisés, Patrice Lopez rapporte ainsi une précision de 64% pour la détection des corrections : Lopez P. (1999) Représenter et utiliser les contraintes de la langue orale à l'aide d'une grammaire lexicalisée d'arbres adjoints. Actes *TALN'99*, Cargèse, France, 445-450.

[102] Dans sa thèse, Zakaria Kurdi a en effet travaillé sur le corpus anglophone TRAINS : Kurdi M.Z. (2003). *op. cit.*

Auteurs	technique	précision	rappel
Bear, Dowding, Shriberg 1992[103]	pattern-matching	50 %	43 %
Dowding *et al*. 1993	pattern-matching	62 %	30 %
Heeman et Allen, 1994-1999[104]	pattern-matching stochastique	87 %	76 %
Core et Schubert 1999[105]	pattern-matching stochastique + méta-règles syntaxiques	47 %	78 %
Kurdi 2002[106]	pattern-matching	88 %	81 %

3.3.3. Première étude de corpus : importance quantitative des réparations

Dans une première étude[107], nous avons cherché à mesurer l'influence du contexte d'interaction sur l'importance quantitative des réparations. Cette étude a été réalisée sur les corpus Murol, Air France et Levelt. Elle montre que la fréquence d'apparition des réparations est corrélée de manière significative avec le degré d'interactivité du dialogue (tableau 2.12).

Tableau 2.12 *— Fréquences d'apparition des réparations sur différents corpus : pourcentage d'énoncés (tours de parole) présentant une répétition ou une correction.*

Corpus	Tâche	Type de corpus	Interactivité	Réparations
Murol	renseignement touristique	H-H simulé	très élevée	34,8 %
Air France	renseignement aérien	H-H réel	assez élevée	9,3 %
Levelt	dessin assisté par ordinateur	magicien d'Oz	assez élevée	13,7 %

D'une manière générale, ces résultats témoignent de l'omniprésence des réparations dans les dialogues oraux. Ces fréquences d'apparition élevées se retrouvent dans d'autres études (tableau 2.13 page suivante) qui ont porté aussi bien sur des corpus pilotes (TRAINS) que sur des dialogues homme-machine réels (VerbMobil). Ces recoupements soulignent, si cela était encore nécessaire, l'importance de cette question pour la communication homme-machine orale.

3.3.4. Seconde étude de corpus : patterns de détection pour le français parlé

Une analyse des usages plus détaillée a été réalisée cette année dans le cadre du stage de DEA d'Iadaloharivola Randria. Elle a porté sur les corpus Air France et OTG, déjà présentés dans ce chapitre. Plus précisément, cette étude a recensé l'ensemble des patterns de réparation présents dans ces corpus. Cette analyse de corpus devait répondre à différentes interrogations :

- l'expression des réparations en français parlé diffère-t-elle de celle de l'anglais, sur lequel se sont concentrées jusqu'ici les approches par patterns ? Pour répondre à cette question, je comparerai par la suite nos observations avec celles d'autres études anglophones,
- existe-t-il des variabilités d'usages qui dépendent du contexte (tâche, type degré d'interactivité, etc.) dans lequel s'inscrit l'interaction ? L'analyse différentielle des résultats obtenus sur les corpus OTG et Air France pourra donner quelques éléments de réponses sur ce point.
- les patterns de réparation recensés nous renseignent-t-ils sur la pertinence de ces approches en

[103] Bear J., Dowding J., Shriberg E. (1992) Integrating multiple knowldege sources for detection and correction of repairs in human-computer dialog. Actes *30th Annual Meeting of the ACL*, *ACL'92*. Newark, Etats-Unis. 56-63.

[104] Heeman P. A., Allen J. F. (1999) Speech repairs, intonational phrases and discourse markers : modelling speakers utterances in spoken dialogue. *Computational Linguistics*, 25(4), 527-573.

[105] Core M. G., Schubert L. K (1999) A syntactic framework for speech repairs and other disruptions, Actes *37th Annual Meeting of the Association for Computational Linguistics, ACL'1999*.

[106] Kudri M.-Z. (2002), Combining pattern matching and shallow parsing techniques for detecting and correcting spoken language extragrammaticalities, Actes *2nd Workshop on RObust Methods in Analysis of Natural language Data*, *ROMAND 2002*, Frascati-Rome, Italie

[107] Letellier-Zarshenas S., Nicolas P., Goulian J., Antoine J.Y. (1999) Inattendus structurels et communication orale finalisée : influence de la tâche et du contexte interactif, Actes *Journées Internationales de Linguistique Appliquée*, *JILA'99*, Nice, France. 176-179.

CHM orale finalisée ? C'est là la question principale à laquelle devait répondre cette étude.

Tableau 2.13 — *Fréquences d'apparition des réparations observées par d'autres auteurs.*

Corpus	Auteurs	Tâche	Type dialogue	Réparations
TRAINS	Heeman et Allen[108]	planification de routage ferroviaire	H-H réel	23 %
VerbMobil	Ruland, Rupp, Spilker, Weber et Worm[109]	traduction médiée par ordinateur (tourisme)	H-M-H réel	20 %

Le tableau 2.14 donne la distribution des réparations suivant la longueur du reparandum. Ce critère, qui accorde un rôle central au reparandum, peut être critiqué d'un point de vue linguistique. Il est incontournable dans la perspective d'une détection automatique à base de patterns.

Tableau 2.14 *— Distribution des réparations suivant la longueur du reparandum (nombre de mots): étude sur les corpus Air France et OTG et comparaison avec d'autres études anglophones.*

Longueur	OTG	Air France	ATIS ARPA (Bear et al,1992)[110]	TRAINS (Heeman,Allen,1999)[111]
1 mot	72,7 %	73,0 %	59,6 %	49,8 %
2 mots	19,4 %	18,7 %	24,4 %	31,5 %
3 mots	6,7 %	6,4 %	8,3 %	12,1 %
4 mots	0,6 %	1,2 %	4,0 %	5,2 %
≥ 5 mots	0,6 %	0,7 %	3,7 %	1,5 %

D'une manière générale, les études présentées dans ce tableau montrent toutes que l'importance des réparations décroît avec la longueur du reparandum. Ce résultat était attendu. Deux conclusions moins intuitives peuvent également être extraites de ces observations.

Tout d'abord, on relève que les réparations des corpus OTG et Air France suivent des distributions très proches[112]. Nous ne détectons donc aucune variabilité d'usage de ce point de vue.

A l'opposé, nos observations diffèrent sensiblement de celles réalisées sur les corpus anglophones. La distribution des réparations est beaucoup plus resserrée sur nos corpus. On observe en particulier une sur-représentation des réparations minimales (un seul mot dans le reparandum), alors que les procédés impliquant quatre mots ou plus sont nettement plus rares que dans les corpus anglophones.

Il serait toutefois risqué de conclure dès à présent à une influence de l'idiome sur ces usages. Ces variations significatives me semblent en effet difficiles à interpréter pour deux raisons :

- d'une part, les corpus étudiés relèvent de contextes d'interaction différents. Nous avons vu (corpus OTG et Air France) que cette dimension ne semble pas constituer une source de variation significative. Cette conclusion demande cependant à être mise à l'épreuve d'autres données pour être confirmée avec certitude,
- d'autre part, il n'est pas certain que les études anglophones répondent exactement à la méthodologie que nous avons suivie. Certes, tous ces travaux suivent le cadre de description

[108] Heeman P. A., Allen J. F. (1999) op. cit.

[109] Ruland T., Rupp C., Spilker J., Weber H., Worm K. (1998) Making the most of multiplicity : a multi-paser multi-strategy for the robust processing of spoken language. Actes *International Conference on Spoken Language Processing. ICSLP'1998*. Sydney, Australie. 570-573.

[110] Bear J., Dowding J., Shriberg E. (1992) *op. cit.*

[111] Les données figurant dans le tableau ne reprennent pas les hésitations, qui sont intégrées parmi les réparations dans l'article originel : Heeman P. A., Allen J. F. (1999) *op. cit*.

[112] Ces travaux n'ont pas encore donné lieu à des analyses statistiques de pertinence. Les résultats présentés dans ce paragraphe me semblent cependant suffisamment significatifs pour que leur validité statistique ne fasse pas de doute.

précis qui a été défini à l'origine par Bear, Dowding et Shriberg[113]. La formalisation des patterns ne peut donc être source d'ambiguïté. Par contre, il existe une latitude non négligeable dans le choix des réparations. Nous avons par exemple décidé de considérer des répétitions d'appui telles que *oui oui oui* ou *non non non*. Ces réparations, très fréquentes en dialogue oral, influent de manière sensible sur la distribution des procédés comme le montre le tableau 2.15. Il serait intéressant de savoir si elles ont été considérées dans les études anglophones citées.

Tableau 2.15 *— Distribution des réparations suivant la longueur du reparandum (nombre de mots), à l'exclusion des répétitions d'appui : étude sur les corpus Air France et OTG.*

Longueur du reparandum	OTG	Air France
1 mot	69,6 %	67,3 %
2 mots	21,7 %	22,7 %
3 mots	7,5 %	7,8 %
4 mots	0,6 %	1,5 %
≥ 5 mots	0,6 %	0,7 %

Le TRAINS Corpus est diffusé librement. C'est pourquoi j'envisage de reprendre prochainement l'étude de Heeman et Allen en suivant notre propre méthodologie, ce qui permettra de s'affranchir d'éventuels biais méthodologiques.

Quoi qu'il en soit, un constat clair peut déjà être tiré de cette étude distributionnelle : en situation de dialogue finalisé, les réparations du français parlé concernent dans leur grande majorité des zones réparées relativement courtes.

Cette apparente simplicité se retrouve dans la distribution des patterns de réparation. Le tableau 2.16 (page suivante) présente les patrons les plus fréquents sur les différents corpus. Plusieurs commentaires peuvent être donnés au vu de ces résultats :

1) les distributions sur les corpus OTG et Air France restent comparables. En particulier, plus de 90% des réparations rencontrées sur ces deux corpus répondent à un des patterns suivants : M1 | M1, R1 | R1, M1 M2 | M1 M2, M1 R2 | M1 R2 et M1 M2 M3 | M1 M2 M3, où | caractérise un point d'interruption non marqué ou un terme d'édition[114]. L'importance relative de ces différents patterns est comparable dans les deux corpus.
2) Les patterns les plus fréquents sont identiques en français et en anglais, à quelques exceptions peu significatives[115].
3) Les distributions rencontrées sur l'anglais diffèrent cependant sensiblement de celles de nos propres études. En particulier, on retrouve un resserrement plus important des distributions sur le français. Une fois encore, il n'est pas aisé d'expliquer ces différences, d'autant plus que les études anglophones ne sont pas toujours cohérentes entre elles. Ainsi, le pattern M1 X^i | M1 (suppression d'éléments du reparandum) représente 8,1% des réparations chez Heeman et Allen. Il n'est cité ni par Kurdi qui a pourtant travaillé sur une partie du corpus, ni par Bear et ses collègues (corpus ATIS). Il ne se retrouve d'ailleurs pas dans nos études.

Par delà ces différences qui demandent à être étudiées plus précisément, cette analyse des usages fournit des renseignements importants sur la réalisation des réparations en français parlé interactif. Une conclusion retient particulièrement mon attention : la répartition des patterns sur les corpus OTG et Air France montre que les réparations sont, dans le cadre du dialogue oral finalisé, moins complexes que ce que pourrait laisser penser la littérature linguistique sur le sujet. C'est ainsi que les enrichissements (ou appauvrissement) lexicaux, qui ont fait l'objet de nombreuses descriptions,

113 Bear J., Dowding J., Shriberg E., Price P. (1993) *op. cit*

114 Ces deux cas de figure ont été regroupés pour permettre une comparaison avec les études anglophones.

115 Le pattern R1| R1 arrive en quatrième position (et non deuxième) sur le corpus ATIS mais reste bien représenté. Le pattern M1 R2 | M1 R2 se classe sixième (et non quatrième) sur le *TRAINS corpus.*

sont marginaux dans nos corpus. Ils n'y représentent environ que 1% des réparations[116]. D'un point de vue quantitatif, il y a donc une différence d'usage sensible entre les monologues ou interviews étudiés par le GARS et les corpus pilotes qui nous concernent.

Tableau 2.16 *— Distribution des patterns des réparations les plus fréquents: étude sur les corpus Air France et OTG et comparaison avec d'autres études anglophones (n.p. : non précisé ; — : pattern non observé sur le corpus). Dans cette étude, le point d'interruption | peut être vide ou marqué par un terme d'édition ET*

CORPUS	OTG	Air France	ATIS (Bear *et al*)	TRAINS (Heeman, Allen)	TRAINS (Kurdi[117])
M1 | M1	48,3 %	56,2 %	15,4 %	26,5 %	54,9 % des répétitions
R1 | R1	24,4 %	15,8 %	8,7 %	14,5 %	31,4 % des corrections
M1 M2 | M1 M2	13,9 %	14,9 %	11,0 %	9,0 %	11 % des répétitions
M1 R2 | M1 R2	4,4 %	1,7 %	10,7 %	3,1 %	8,6 % des corrections
M1 X^i | M1	—	0,4 %	< 3,9%	8,1 %	n.p. (< 3 %)
M1 | X^i M1	—	0,4 %	6,7 %	1,8 %	n.p. (< 3 %)
R1 X^i | R1	—	—	< 3,9%	2,1 %	n.p. (< 3 %)
R1 M2 | R1 M2	—	0,9 %	3,9 %	2,1 %	6,7 % des corrections
M1 M2 X^i | M1 M2	—	0,1 %	< 3,9%	3,7 %	n.p. (< 3 %)
M1 M2 | M1 X^i M2	—	0,1 %	7,5 %	0,6 %	2 %
M1 M2 R3 | M1 M2 R3	2,8 %	1,0 %	n.p.	1,1 %	6,7 % des corrections
M1 M2 M3 | M1 M2 M3	3,9 %	3,1 %	n.p.	2,3 %	3,5 % des répétitions
Total 5 patterns principaux	**94,9 %**	**91,7 %**	**53,3 %**	**61,8 %**	**62 %**

Cette constatation ne doit pas laisser croire à une excessive simplicité des réparations. Nous avons vu avec l'énoncé 2.4 (cf. § 3.3.2) que le travail de reformulation pouvait concerner plusieurs réparations successives, qui correspondent aux différents entassements paradigmatiques de l'analyse en grille. Je constate simplement que ces réparations individuelles suivent des schémas relativement simples de répétition (patterns $M^i \mid M^i$) ou de remplacement (patterns R1 | R1 ou M1 R2 | M1 R2).

Ce résultat n'était pas attendu. Il autorise un certain optimisme sur la mise en œuvre de méthodes de détection par patterns efficaces et couvrantes. Bien entendu, cette analyse des usages ne saurait répondre aux questions d'ordre technologique qui se posent désormais à nous :

- comment limiter le caractère sur-génératif de certains patterns,
- comment gérer la succession, voire l'imbrication, des réparations,
- comment distinguer les répétitions des corrections.

Je ne vais pas détailler dans ce chapitre consacré à la linguistique de corpus les solutions que nous sommes en train d'étudier en réponse à ces problèmes. Ce que je retiendrai ici, c'est qu'une étude de corpus pilote a pu montrer qu'une approche par patterns était, à l'encontre de nos prévisions, compatible avec les usages langagiers de la CHM orale. Les patterns qui ont été caractérisés par

116 Patterns présentant un X dans le reparandum (appauvrissement) ou l'altération (enrichissement).

117 Kudri M.-Z. (2002) *op. cit*.

cette analyse d'usages ont également permis le prototypage d'un premier système de détection. Il seront prochainement affinés sur un corpus de magicien d'Oz qui sera réalisé dans la cadre du projet MEDIA de l'appel d'offre TECHNOLANGUE (cf. chapitre 3 § 3).

3.4. Co-référence anaphorique et traitements linguistiques de surface

La dernière étude que je vais présenter a été amorcée dans le cadre du stage de DEA de Julien Foulon. Je l'ai ensuite approfondie et étendue en variant les corpus d'observation considérés. Cette analyse de corpus constitue une illustration d'une des caractéristiques de ma recherche, à savoir la tentative d'adaptation de méthodes issues du TALN (langage écrit) à la problématique du traitement du langage parlé. Plus précisément, cette étude prospective s'intéresse à la résolution des anaphores pronominales dans le cadre du dialogue oral homme-machine. A l'opposé de l'étude précédente sur les réparations, nous verrons que l'analyse de corpus pilote réalisée conclut plutôt à l'inadéquation des techniques envisagées.

3.4.1. Anaphore et dialogue homme-machine : méthodes génériques pour une résolution

Jusqu'à présent, les théories et méthodes de résolution des co-références anaphoriques se sont surtout intéressées aux textes écrits. Cette problématique a ainsi fait l'objet de nombreuses études en recherche d'information textuelle, même si ces travaux ont généralement une portée assez limitée (pronoms de la troisième personne).

A l'opposé, si le calcul de la référence est essentiel en dialogue homme-machine[118], la résolution des seules co-références pronominales est facilitée par le caractère finalisé de l'interaction. C'est pourquoi de nombreux systèmes de dialogue oral se contentent de résoudre ce problème à l'aide de techniques *ad hoc* de consultation de l'historique du dialogue et de la tâche.

Dans la perspective d'une communication orale homme-machine générique, il me semble cependant intéressant d'appliquer à l'oral des méthodes de résolution plus générales. C'est l'objectif de cette étude de corpus qui cherche à vérifier si certaines méthodes développées pour des textes écrits peuvent s'appliquer au dialogue oral.

3.4.2. Anaphore et ingénierie des langues : méthodes robustes de résolution des co-références

Avec son « algorithme naïf », Hobbs a montré dès 1978 qu'un ensemble de règles assez simple pouvait permettre de résoudre plus de 88% des anaphores pronominales sur un texte technique[119]. En première approximation, cet algorithme se fonde sur des heuristiques relativement intuitives. D'une part, il cherche prioritairement les antécédents dans l'énoncé courant avant d'aller étudier la phrase précédente[120]. Ensuite, l'algorithme recherche les antécédents compatibles en genre et nombre avec le pronom, dans une stratégie de recherche gauche droite et en largeur d'abord (préférence aux constituants immédiats plutôt qu'aux sous-constituants enchâssés). Le parcours gauche droite est justifié par un postulat de préférence de rattachement au sujet puis à l'objet sur les autres compléments. Cette stratégie s'applique donc aux langages SVO et constitue de ce point de vue une claire illustration du caractère anglo-centré des méthodes en ingénierie des langues.

S'il a connu plusieurs tentatives d'amélioration[121], cet algorithme reste une référence en la matière. La recherche des différents antécédents possibles repose par contre implicitement sur un parcours orienté de la structure syntaxique de l'énoncé. Cette approche pose deux problèmes dans la perspective d'une utilisation sur la parole spontanée :

- d'une part, les réparations, incises et autres procédés de l'oral spontané cassent fréquemment la

[118] Pierrel J-M., Romary L. (2000) Dialogue homme - machine. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 331-350 (paragraphe 15.3 : pp. 337-343).

[119] Hobbs J. R. (1978) Resolving pronoun references. *Lingua*, 44, 331-338.

[120] Hobbs avait remarqué dans une étude de corpus que 98 % des antécédents des pronoms non déictiques se trouvaient dans l'une de ces deux phrases.

[121] Ces améliorations se sont traduites par des gains d'efficacité relativement marginaux. Ainsi, l'algorithme RAP de Lappin et Leass, présente un robustesse de 86 % sur un texte où Hobbs aurait obtenu 82% de réussite : Lappin S., Leass H.J. (1994) An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4), 535-561.

structure de ces énoncés, ou du moins rendent difficile leur analyse automatique.

- d'autre part, la forte variabilité faible que nous avons observée sur nos corpus pilotes (cf. § 3.2) risque de mettre à mal une stratégie d'analyse gauche droite, même si nous avons vu que le français parlé restait majoritairement un langage de type SVO.

Cette dépendance à une analyse syntaxique préalable a été relevée par d'autres auteurs[122]. Même si elle demanderait à être étudiée de plus près, cette limitation m'a conduit à rechercher des solutions alternatives dans les travaux les plus récents du domaine. Globalement, les recherches sur la référence ont pris deux directions opposées au cours de la dernière décennie :

- d'un côté, certains chercheurs ont mis en œuvre des modèles théoriques qui permettent une description beaucoup fine de la référence mais qui n'ont pas encore donné lieu à des réalisations opérationnelles. C'est le cas par exemple de la théorie du centrage (*centering theory*) qui propose un modèle local fondé sur les deux derniers énoncés[123].
- d'un autre côté, certaines recherches se sont focalisées sur des questions d'efficacité et de robustesse dans une démarche proche de celle du TAL robuste (*shallow parsing*). Certains chercheurs ont ainsi proposé des approches assez grossières mais exigeant un minimum de connaissances.

L'algorithme développé par Mitkov constitue l'archétype de cette seconde démarche[124]. A la suite d'une segmentation en chunks de l'énoncé, Mitkov considère comme antécédents possibles tous les groupes nominaux des deux énoncés précédents qui s'accordent en genre et nombre avec le pronom. L'algorithme procède ensuite à un classement des candidats suivant des règles de préférence *ad hoc* qui sont le plus souvent non justifiées par des considérations linguistiques (désavantage des groupes indéfinis sur les définis, désavantage des groupe nominaux introduisant un groupe prépositionnel, voire des règles plus spécifiques).

Ce type d'approche laisse une impression mitigée. D'un côté, il répond à une démarche purement ingénierique dont on perçoit aisément les limites en terme de finesse d'analyse. De l'autre, il s'intègre dans une démarche de type TAL robuste qui peut constituer un bon compromis entre recherche d'efficacité et de pertinence linguistique. L'algorithme développé par Mitkov est d'ailleurs celui qui présente à l'heure actuelle les meilleurs résultats (89 % de réussite là où Hobbs aurait fait 82%) dans un cadre certes limité.

C'est précisément pour avoir un avis plus clair sur l'intérêt de ces différentes approches que j'ai lancé une étude sur corpus pilote dont l'objectif était de vérifier dans quelle mesure les heuristiques les plus utilisées dans le domaine (accord en genre et nombre, préférence pour le sujet etc.) étaient pertinentes en dialogue oral.

3.4.3. Etude de corpus : méthodologie

Cette analyse des usages a concerné le corpus OTG (DEA Julien Foulon) ainsi que les corpus Air France et Murol qui ont déjà été présentés plus haut. Nous avons considéré tous les pronoms personnels sujets, conjoints ou disjoints de la troisième personne. Au total, plus de 500 exemples d'anaphores pronominales ont ainsi été relevés. Chaque occurrence a fait l'objet d'une annotation détaillée suivant différentes caractéristiques :

- présence ou non d'anaphore. Les pronoms personnels de la troisième personne ne réfèrent en effet pas toujours à un élément du discours (tournures impersonnelles, références indirectes).
- distance entre le pronom et son antécédent, mais également de la dernière référence à

[122] Kennedy C., Boguraev B. (1996) Anaphora for everyone : pronominal anaphora resolution without a parser. Actes *16th International Conference on Computational Linguistics, COLING-96*, Copenhague, Danemark. 113-118.

[123] Grosz B.J., Joshi A.K., Weinstein S. (1995) Centering : a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2), 203-225.

[124] Mitkov R. (1998) Robust pronoun resolution with limited knowledge. Actes *36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, *ACL-COLING'98*, Montréal, Canada, 869-875.

l'antécédent, dans le cas d'anaphores multiples. Cette distance est mesurée en nombre de tours de parole. Cette métrique est le pendant des distances exprimées en nombre de phrases dans les travaux sur l'écrit. Les cataphores sont également décrites par cet attribut.

- fonction syntaxique (*prédicat*, *sujet*, *objet*, autres compléments, modifieurs) ainsi que le genre (masculin, féminin, indéterminé) et nombre (*singulier*, *pluriel*, *indéterminé*) du pronom, de son antécédent et éventuellement de la dernière référence à cet antécédent,

Nous n'avons pas considéré le caractère défini ou indéfini des éléments concernés. Dans la perspective d'une analyse à la Mitkov, ce caractère ne peut être étudié qu'à un niveau syntaxique. Or, l'observation de corpus nous a montré la fragilité d'une caractérisation purement grammaticale du caractère défini des objets auxquels réfèrent les énoncés.

3.4.4. Fréquence d'apparition et distribution des anaphores pronominales

A ce jour, cette étude distributionnelle n'a été menée que sur les corpus Air France et Murol. Le recours à l'anaphore peut être relativement variable d'une situation interactive à une autre. Ces deux corpus semblent cependant assez homogènes de ce point de vue, puisque l'on n'y relève pas de différence significative dans la fréquence d'apparition des anaphores (tableau 2.17). Les études que je mène sur d'autres corpus devraient permettre de mieux appréhender cette faible diversité.

Tableau 2.17 *— Fréquence d'apparition des anaphores pronominales (3° personne)*

Corpus	% tours de parole avec anaphore	nombre d'anaphores par mot
Air France	4,5	7 ‰
Murol	5,9	6 ‰

Plus intéressante est l'analyse de la distribution des pronoms suivant le phénomène mis en jeu (tableau 2.18 page suivante). Pour chaque occurrence, nous avons distingué trois situations :

- anaphore (ou cataphore) directe ou indirecte.
- anaphore implicite, forme complexe d'anaphore inférentielle pour laquelle il n'existe aucune trace lexicale dans le dialogue permettant d'amorcer le calcul de la référence :

  (2.9) *à 12h10 attendez je* ***l****'appelle sur une autre ligne* (AF I-13 :C7)

  Dans cet exemple, il n'a jamais été fait (et ne sera jamais fait) mention explicite au cours du dialogue du client, qui est visiblement l'objet de la référence.
- usage impersonnel figé (*il y a*) ou non figé : *il est certain qu'il y a* (AF I-64 :O10), marque de certitude générale à comparer à : *Jean*, *il est certain qu'il y a eu de la vie sur Mars*.

Nous ne nous intéresserons pas ici aux impersonnels figés, qui peuvent être aisément détectés par l'analyse automatique. Le tableau 2.18 donne la distribution des pronoms dans les autres situations.

Tableau 2.18 *— Distribution des pronoms suivant leurs usages (corpus Air France et Murol)*

Corpus	Anaphores / cataphores	anaphores implicites	Impersonnels non figés
Air France	80 %	12 %	8 %
Murol	91 %	3 %	6 %

Si on constate sur les deux corpus une prédominance des anaphores directes ou indirectes, deux remarques peuvent être faites. D'une part, on observe une sur-représentation des anaphores implicites sur le corpus Air France. Ce résultat traduit l'influence du contexte applicatif sur la référence. On observe en effet qu'une part importante de ces observations correspond à une mention implicite du client, comme dans l'exemple (2.9). Du fait du caractère extrêmement ciblé de la tâche (réserver un billet pour un client), les inférences nécessaires à la résolution des références implicites sont réduites, d'où un recours facilité à ce procédé.

D'autre part, cette étude montre que dans 9% à 20 % des cas, le pronom ne réfère pas ou n'a pas d'antécédent exprimé dans le dialogue. Les méthodes que j'ai évoquées plus haut (§ 3.4.2) ne disposent apparemment d'aucun mécanisme pour caractériser ces usages particuliers. Ils devraient donc systématiquement générer des erreurs dans ces situations. Il s'agit d'une limitation importante à leur application en dialogue oral homme-machine.

3.4.5. Le postulat de l'accord en genre et en nombre

Le tableau 2.19 fait la synthèse du pourcentage d'accord en genre et en nombre entre, d'une part, le pronom et son antécédent, et d'autre part le pronom et la dernière référence à cet antécédent

Tableau 2.19 *— Respect de l'accord en genre et en nombre entre le pronom et son antécédent*

Corpus	Accord pronom / antécédent		accord pronom / dernier référent	
	genre	nombre	genre	Nombre
OTG	78 %	50 %	—	—
Air France	94 %	82 %	96 %	91 %
Murol	100 %	97 %	100 %	99 %

Si, dans le cas du corpus Murol, le respect de l'accord est confirmé, cette règle est déjà moins partagée dans le cas du corpus Air France, pour devenir non opérante sur le corpus OTG.

Sur les trois corpus, l'accord en nombre est moins respecté que l'accord en genre. Cette observation peut étonner, lorsqu'on connaît la fragilité du genre en français. Cette situation s'explique pourtant aisément. Le plus souvent, l'absence d'accord correspond en effet à des emploi métonymique du pronom *ils* qui entraîne essentiellement des problèmes de respect du nombre. Les exemples (2.10) et (2.11) sont représentatifs de ce type d'usage :

(2.10) *ce qu'il faudrait faire pour ça c'est contacter* ***l'OMS*** *c'est l'office municipal des sports* (OTG-3ap0038 :016)
…
ils *ouvrent qu'à 15 heures hein.* (OTG-3ap0038 :020)

(2.11) *voilà e pouvez vous me dire s'il y a eu une réponse de* ***l'hôtel*** (AF I-89 :C5)
…
*oui mais // mais est ce est ce qu'****ils*** *vous ont déjà répondu* (AF I-89 :C8)

Ce genre d'anaphore indirecte entre une classe et ses éléments particuliers ou constitutifs a déjà été décrite en linguistique[125] mais n'a pas fait l'objet, à ma connaissance, d'études distributionnelles. Elles remettent pourtant en question — en dialogue oral finalisé tout au moins — un postulat d'accord sur lequel reposent aveuglément les techniques de résolution des anaphores pronominales.

Notons enfin que l'accord est un peu plus respecté dans le cas d'une mise en correspondance entre le pronom est la dernière référence à l'antécédent (tableau 2.19). Une fois la métonymie mise en place (et acceptée par l'interlocuteur), rien ne s'oppose à la poursuite de cet usage. Par exemple, le dialogue de l'exemple (2.11) se poursuit comme suit :

(2.12) […] *si on ne* ***les*** *relance pas* […] (AF I-89 :C5)

La variabilité des observations entre les corpus semble traduire l'influence du contexte d'interaction sur la réalisation de ces anaphores : suivant les concepts qui constituent l'univers de la tâche, un cadre applicatif peut en effet être plus favorable qu'un autre à la mise en place de ces métonymies. Ces observations montrent en tous cas que l'adaptation à la CHM orale d'indices largement utilisés

[125] Charolles M. (2002) La référence et les expressions référentielles en français. Ophrys, Gap, France ; Berrendonner A., Reichler-Béguelin M.J. (1995) Accords associatifs, *Cahiers de praxématique*, 24, 14-21.

sur des textes écrits n'a rien d'évident

3.4.6. La préférence accordée aux antécédents les plus proches

Un autre postulat, utilisé par la plupart des techniques de résolution, semble également plus friable en dialogue oral finalisé. Il s'agit de la préférence de rattachement vers les antécédents candidats les plus proches. Le tableau 2.20 donne la distribution des distances en nombre de tours de parole, entre le pronom et son antécédent (ou également avec la dernière référence à ce dernier).

Tableau 2.20 *— Distribution des anaphores pronominales suivant la distance en nombre de tours de parole entre le pronom et l'antécédent (même tour de : dist. = 0 ; tour précédent: dist. = 1)*

Corpus	Distance pronom / antécédent				distance pronom / dernier référent			
	0	**1**	**2**	**> 2**	**0**	**1**	**2**	**>2**
OTG	42%	22%	22%	**13%**	—	—	—	—
Air France	35%	15%	17%	**33%**	46%	19%	19%	**16%**
Murol	25%	18%	32%	**25%**	27%	24%	33%	**16 %**

Ces résultats ne recoupent que très partiellement les observations effectuées sur des textes écrits. Hobbs a montré dans ses études sur des textes techniques que 90% des anaphores pronominales sont interphrastiques, et que 98% des antécédents se situent dans la phrase courante ou la précédente. En dialogue oral, il paraît cohérent de transposer cette caractérisation à l'énoncé courant et au tour de parole précédent de chacun des interlocuteurs (distances 0, 1 ou 2 dans le tableau 2.20).

D'après nos observations, une part notable (13% à 33% suivant le corpus) des antécédents appartiennent pourtant à un tour de parole antérieur. De par son ancrage interactif, le dialogue oral semble ainsi pouvoir référer à des éléments beaucoup plus lointains que le langage écrit.

On notera par ailleurs que lorsqu'il réalise une anaphore pronominale, le locuteur réfère en majorité à ses propres productions (tableau 2.21). Cette stratégie se trouve même légèrement renforcée si le locuteur a déjà co-référé lui-même à cet élément. Les résultats sont ici d'une remarquable stabilité.

Tableau 2.21 *— Distribution des anaphores pronominales suivant le locuteur qui a exprimé l'antécédent ou la dernière référence à cet antécédent*

Corpus	Pronom / antécédent		pronom / dernier référent	
	même locuteur	**Interlocuteur**	**même locuteur**	**Interlocuteur**
OTG	69 %	31 %	—	—
Air France	70 %	30 %	75 %	25 %
Murol	67 %	33 %	68 %	32 %

Si la récence d'information a une influence réelle sur la détermination des antécédents, cette distribution semble cependant moins resserrée à l'oral qu'à l'écrit. En dialogue oral, ce type d'influence ne peut donc revêtir un caractère sélectif aussi marqué qu'à l'écrit.

3.4.7. Le manque de fiabilité des critères syntaxiques pour la résolution des références

En conclusion, cette analyse des usages tend à montrer que les anaphores pronominales ne suivent pas la même logique en dialogue oral finalisé et dans les textes écrits. Des indices grammaticaux tels que l'accord en genre et en nombre, ou la distance entre le pronom et son référent, permettent d'atteindre des taux de robustesse de l'ordre de 90% sur des textes écrits. Nos études montrent que ces indices sont nettement moins fiables sur l'oral finalisé.

Cette étude préliminaire doit bien entendu être complétée et affinée avant de conclure à l'impossibilité d'adapter à la CHM orale les méthodes de résolution développées en TAL robuste.

Mon objectif reste de résoudre, avec des méthodes de bas niveau donc génériques, au moins une partie des co-références anaphoriques. Pour l'heure, cette analyse des usages semble confirmer par défaut la pertinence des traitements pragmatiques utilisés jusqu'ici en dialogue oral.

Comme pour l'ensemble des études présentées dans ce chapitre, on notera que cette conclusion n'a nécessité le développement d'aucun système complexe. C'est là un intérêt non négligeable de ce type d'étude amont.

4. CONCLUSION

Les analyses d'usages que je viens de présenter demandent à être poursuivies sur d'autres domaines pour atteindre une représentativité encore plus grande. Dans le cadre du programme *Parole Publique*, nous disposerons dès la fin 2004 d'une banque de corpus qui pourra répondre plus amplement à nos besoins en matière d'analyse prospective.

D'une manière générale, ce type d'analyse demande un effort important de recherche qui explique certainement qu'elles soient peu prisées en ingénierie des langues. J'espère cependant avoir réussi à montrer sur quelques exemples l'intérêt de tels travaux pour la conduite des recherches en communication homme-machine et plus généralement en ingénierie des langues. L'impact de ces travaux serait d'ailleurs renforcé s'il l'on disposait d'une banque de corpus de dialogue oral normalisée, qui permettrait de mener des études variationnelles fines suivant différentes caractéristiques[126]. C'est aussi l'objectif que vise, à son échelle, le programme *Parole Publique*.

La morale que l'on pourrait tirer de ces travaux pourrait ainsi être la suivante : on analyse bien que ce que l'on connaît bien. Si l'ingénierie des langues travaille désormais sur des données réelles d'envergure, les caractéristiques linguistiques de ce matériau retiennent malheureusement trop rarement l'attention. Comme nous allons maintenant le voir, cette méconnaissance se retrouve à l'aval de la conception des technologies langagières. Les méthodologies actuelles d'évaluation font en effet peu cas du fait linguistique.

[126] Pour se faire une idée (sur d'autres thématiques de recherche) de l'apport que l'on pourrait tirer d'une étude variationniste de la langue en situation de dialogue oral, on pourra consulter le numéro spécial de la revue *Langages* consacré aux travaux de William Labov : Gadet F. (Dir.) (1992) Hétérogénéité et variation : Labov, un bilan. *Langages*, 108, Larousse, Paris, France.

3. Quelle évaluation pour l'ingénierie des langues ?

Juger est quelquefois un plaisir,
Comprendre en est toujours un

Henri de Régnier, *Donc…*

Si on devait définir par un mot le renouvellement des pratiques en traitement automatique des langues au cours des quinze dernières années, c'est sans aucun doute le terme « évaluation » qui retiendrait l'esprit. La mise en place de campagnes de tests d'envergure — aux Etats-Unis puis en Europe — a en effet initié l'aggiornamento qui a conduit le TALN à s'attaquer à des données et des problèmes réels. Aujourd'hui, toute recherche en ingénierie des langues donne lieu à une validation expérimentale sérieuse, si possible dans le cadre de programmes confrontant plusieurs systèmes. Ces campagnes de tests fournissent une photographie de l'état de l'art qui nous renseigne sur le chemin parcouru et sur la distance qui nous sépare d'applications destinées au grand public.

Ce recours généralisé à l'évaluation a conduit le traitement automatique des langues à repenser ses objectifs, ses pratiques voire ses approches théoriques. A l'opposé, cette réflexion a peu concerné l'évaluation en elle-même. De multiples paradigmes de test ont certes été proposés. Ils se situent cependant majoritairement dans le même cadre épistémologique. C'est-à-dire qu'ils suivent une démarche purement ingénierique reposant sur des métriques bien maîtrisées mais qui ne nous renseignent qu'imparfaitement sur la conduite de recherches futures.

Après un état de l'art critique des pratiques de test en ingénierie des langues, je ferai plusieurs propositions pour répondre à ces insuffisances. Je montrerai en particulier l'intérêt d'une évaluation objective de type diagnostic. Ces propositions seront détaillées dans le cadre de la compréhension de la parole pour laquelle j'ai défini, avec d'autres collègues, deux paradigmes de test novateurs qui ont donné lieu à validation expérimentale.

1. EVALUER : POUR QUOI, POUR QUI ET COMMENT ?

L'évaluation des systèmes n'étant pas une fin en soi, il importe de connaître les buts que l'on poursuit lors du la mise en place d'une campagne de test. Tout d'abord, il faut se demander à qui s'adresse l'évaluation. On distingue ainsi traditionnellement deux types d'évaluation :

- une **évaluation objective** poursuit un objectif purement technologique. Elle s'adresse aux concepteurs et doit fournir une mesure des performances du système en dehors de toute considération sur sa perception par les utilisateurs. Elle peut donner lieu à une comparaison entre systèmes et présenter un pouvoir diagnostic plus ou moins élevé.
- une **évaluation subjective** est fondée au contraire sur le jugement des utilisateurs finaux. On cherche ici à mesurer l'adéquation du système à la tâche ainsi que son utilisabilité en termes de convivialité, fiabilité et facilité d'utilisation. L'évaluation repose sur des indices mesurant la satisfaction des utilisateurs suivant différents points de vue. Cette satisfaction est estimée par des entretiens post-expérimentation ou plus fréquemment par l'analyse de questionnaires[1] détaillés remplis par les sujets (tableau 3.1 page suivante).

1.1. Evaluation subjective : un déficit d'analyse des facteurs humains

Dans la perspective d'une interaction conviviale entre l'utilisateur et la machine, l'évaluation subjective des systèmes interactifs devrait être au centre de nos pratiques. En dépit de certaines expérimentations ou réflexions (EAGLES, DISC), force est de constater qu'il n'en est rien. Comme le regrettent Dybkjaer et Bersen, il existe peu de recherches sur l'utilisabilité de ces systèmes et plus

[1] Hone K.S., Graham R. (2000) Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, 6 (3-4). 287-303.

généralement sur les aspects humains du dialogue homme - machine[2]. Le constat que faisait Margaret King en 1995 est donc toujours d'actualité[3] :

> *« It is quite astonishing how little attention is paid to users in the published literature on evaluation »*

Plusieurs causes président à cette situation. En premier lieu, l'expérimentation des systèmes en situation réelle constitue une activité lourde et coûteuse. Elle ne peut donc être réalisée que dans le cadre de projets d'envergure.

Tableau 3.1 *— Evaluation subjective des systèmes de dialogue oral : exemple de questionnaire utilisée pour le système de ARISE du LIMSI[4].*

Questions du questionnaire	**Fonctionnalités étudiées**
Le système était facile à comprendre (oui/non)	Synthèse de parole
Le système à compris ce que j'ai dit (oui/non)	reconnaissance et compréhension
J'ai obtenu l'information que je demandais (oui/non)	compréhension contextuelle
La vitesse des échanges était appropriée (oui/non)	vitesse
A tout moment je savais quoi dire (oui/non)	aide à l'utilisateur
Le système a clairement exprimé ce qu'il comprenait (oui/non)	génération de réponse
Les suggestions ou questions du système m'ont aidé (oui/non)	stratégies de génération

Par ailleurs, de nombreux facteurs humains sont susceptibles de biaiser les conclusions d'une évaluation subjective. On citera par exemple l'extrême variabilité de l'attitude des sujets face à la machine, le problème de leur représentativité au regard de la population à laquelle le système est destiné, ou encore le manque d'investissement des sujets dans l'expérience. L'objectif étant d'étudier le comportement naturel des utilisateurs, il n'est pas possible de contrôler ces facteurs humains comme le fait la psychologie expérimentale. L'augmentation du nombre de sujets testés peut limiter les risques de biais. Elle ne fait malheureusement qu'accroître l'effort nécessaire à la mise en place de ces expérimentations.

Ces problèmes méthodologiques peuvent difficilement être contournés. D'autres facteurs résultent au contraire de nos pratiques quotidiennes de conception. En particulier, je pense que le recours limité à l'évaluation subjective provient des difficultés que l'on rencontre à relier métriques subjectives et considérations techniques[5]. D'une part parce que l'opinion de l'utilisateur est un jugement global qui se traduit rarement en terme de fonctionnalités précises, mais également parce qu'un serveur interactif est un système complexe qui met en jeu de multiples traitements aux comportements parfois antagonistes. C'est ainsi que l'on arrive parfois à des résultats contradictoires entre l'amélioration (objective) des performances d'un module particulier et le jugement subjectif qu'ont les utilisateurs de cette évolution[6].

---

2 Dybkjaer L., Bersen N.-O. (2000) Usability issues in spoken dialogue systems. *Natural Language Engineering*, 6 (3-4), 243-271 (paragraphe I).

3 King M. (1995) Human factors and user acceptability. In Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon, USA. Disponible sur la Toile : http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. 491-494.

4 Exemple tiré de : Rosset S. (2000) Stratégies et gestionnaire de dialogue pour des systèmes d'interrogation de bases de données à reconnaissance vocale. Doctorat Université Paris XI, Orsay, France. publié comme rapport de recherche 2001-18 du LIMSI-CNRS, Orsay, France. septembre 2001 (tableau page 234).

5 Ce problème n'est pas propre au dialogue oral homme-machine. Voir par exemple les difficultés qu'il peut exister à relier qualités objectives et subjectives des résumés automatiques de texte : Minel J.-L., Nugier S., Piat G. (1997) Comment apprécier la qualité des "résumés" automatiques de textes ? Les exemples des protocoles FAN et MLUCE et leurs résultats sur SERAPHIN. Actes *1ères Journées Scientifiques et Techniques du réseau FRANCIL*, *JST-FRANCIL'97*, Avignon, France. 227-232

6 Danieli M., Gerbino E. (1995) Metrics for evaluating dialogue strategies in a spoken language system. Actes *AAAI Spring symposium on empirical methods in discourse interpretation and generation*. Standford, CA. 34-39.

Le paradigme PARADISE[7] cherche à appréhender cette relation entre perception subjective et comportement objectif du système. Formellement parlant, PARADISE réalise une analyse multicritère qui associe par régression linéaire multiple des résultats issus de métriques différentes. Son intérêt est d'utiliser cette régression pour établir des corrélations entre différentes mesures objectives et un critère subjectif (la satisfaction de l'utilisateur) que l'on cherche à optimiser. Marylin Walker et ses collègues proposent ainsi d'utiliser une hiérarchie d'indices objectifs qui concernent aussi bien la reconnaissance, la compréhension que le dialogue[8]. Cette méthodologie n'est cependant pas liée à une typologie de mesures particulière. Au final, on obtient une estimation de la satisfaction de l'utilisateur en fonction (combinaison linéaire) des indices objectifs.

Le paradigme PARADISE a été utilisé dans le récent programme d'évaluation Communicator[9] de la DARPA américaine. Ce projet concerne l'évaluation des systèmes interactifs dans le domaine du renseignement touristique. Il a montré que quatre mesures objectives principales (taux d'erreur de la reconnaissance de parole, complétude par rapport à la tâche, durée de l'interaction et nombre d'échanges moyens par dialogue) permettent d'expliquer 37% de la variance de la satisfaction des utilisateurs. Le système ARISE (renseignement ferroviaire) du LIMSI a également évalué à l'aide de ce paradigme[10]. Dans ce cas, la satisfaction de l'utilisateur a été expliquée à hauteur de 44% de la variance par trois mesures objectives : les taux d'erreurs de l'interprétation contextuelle et de la reconnaissance de la parole ainsi que le taux de répétitions des utilisateurs.

Nous manquons de recul pour juger de l'apport réel de PARADISE. On peut ainsi s'interroger sur le degré de généricité des fonctions de corrélations obtenues. Sont-elles propres à une application ou conservent-elles une certaine pertinence sur d'autres domaines ? Par ailleurs, ce paradigme s'appuie sur des métriques objectives globales qui ne nous renseignent pas sur les améliorations précises à apporter au système. Cette limitation sera précisément discutée au cours du paragraphe suivant.

Quels que soient les mérites du paradigme PARADISE, je vais tenter de montrer dans ce chapitre qu'il est possible de concilier d'une autre manière évaluation objective et prise en compte des besoins des utilisateurs. Ma proposition, qui s'appuie sur une analyse préalable des usages langagiers, implique une remise en question des pratiques de test en ingénierie des langues. Le paragraphe suivant vise précisément à montrer en quoi les évaluations objectives actuelles ne répondent que partiellement aux questions et enjeux de l'ingénierie des langues.

1.2. Evaluation objective : limites d'une évaluation purement technologique

On peut parler d'évaluation objective lorsque la validation d'un système repose sur l'estimation de mesures de performances pouvant être reproduites sans modification des résultats. Cette définition très générale peut recouvrir des pratiques de tests très différentes. Dans un texte de synthèse sur l'évaluation, Lynette Hirschman et Henry Thompson opèrent une distinction intéressante entre deux types d'évaluation objective[11] :

- l'**évaluation diagnostic** cherche à caractériser finement le comportement du système. Elle repose sur la définition de séries de tests répondant à une taxonomie de situations particulières,

---

7 Walker M., Kamm C., Litman D. (2000) Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6 (3-4), 363-377 ; Walker M., Kamm C., Boland J. (2000) Developing and testing general models of spoken dialogue system performance, Actes *2nd International Conference on Language Resources and Evaluation*, *LREC'2000*, Athènes, Grèce. 189-196.

8 Les auteurs distinguent d'une part des mesures d'efficacité brutes et d'autres part des mesures dites qualitatives, qui reposent néanmoins toujours sur une observation des performances globales des composants du système interactif.

9 Walker M., Passonneau R., Boland J. (2001) Quantitative and Qualitative Evaluation of DARPA Communicator Spoken Dialog Systems, Actes *ACL/EACL'2001*. Toulouse, France.

10 Bonneau-Maynard H., Devillers L., Rosset S. (2000) Predictive performance of dialog systems. Actes *2nd International Conference on Language Resources and Evaluation*, *LREC'2000*, Athènes, Grèce.

11 Hirschman L., Thomspson H. S. (1995) Overview of evaluation in speech and natural language processing. In Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon, Etats-Unis. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. 475-481.

- l'**évaluation des performances** mesure de manière globale les performances du système sur une tâche particulière. On utilise dans ce cas des mesures très générales telles que des taux de rappel et de précision[12], robustesse[13] (*accuracy*) ou encore des mesures statistiques de perplexité.

Les campagnes d'évaluation qui ont marqué l'évolution de l'ingénierie des langues ont clairement privilégié l'évaluation des performances à l'évaluation diagnostic. On retrouve ici une orientation technologique de l'ingénierie des langues qui conduit à la recherche d'estimateurs simples et fiables. Comme le rappelle Margaret King, ces campagnes d'évaluation sont faites par des technologues (les concepteurs de systèmes) pour des institutions ayant avant tout des motivations technologiques qui ne laissent que peu de place aux facteurs humains ou linguistiques[14].

L'apport de l'évaluation des performances n'est bien entendu plus à démontrer. Elle fournit en effet une photographie instructive des performances brutes des systèmes à un instant donné. Compte tenu de leur caractère global[15], les campagnes d'évaluation des performances présentent toutefois un pouvoir prédictif limité. Il est en effet difficile d'interpréter les performances générales d'un système à partir de leurs conclusions : sont-elles dues à une bonne modélisation du langage, à la mise en place d'une stratégie d'analyse robuste, ou plus simplement à une excellente adaptation à la tâche ? Une évaluation globale ne peut répondre à ces questions, de même qu'elle ne peut caractériser finement les classes de problèmes qu'un système ne sait résoudre correctement.

En conséquence, l'évaluation globale des performances est peu utile à la conduite des recherches en ingénierie des langues. Comme le rappellent Caelen, Zeiliger, Siroux et leurs collègues[16] :

> *« […] on a autant besoin de critères **prédictifs** que de critères de **performances** pour l'évaluation prospective des systèmes. Les critères **prédictifs** sont même à certains égards presque plus utiles dans la phase de conception, dans la mesure où ils évitent de s'engager dans des méthodes qui n'ont aucune chance de succès »* (souligné par les auteurs)

Cette insuffisance est bien connue des chercheurs, qui ont fréquemment recours à l'étude de sessions d'utilisation (analyse de *fichiers de log* ou *logfiles*) pour détecter les dysfonctionnements de leur système sur des situations précises. Dans ce cas, nous sommes en présence d'une évaluation diagnostique au sens de Hirschman et Thompson. Ces pratiques individuelles ne sont toutefois ni systématiques ni généralisables. Elles sont donc inadaptées à des campagnes d'évaluation regroupant plusieurs participants. Elles suggèrent en revanche que le caractère prédictif de l'évaluation va de pair avec son caractère discriminant. Je reviendrai sur cette observation plus loin.

---

12 Les taux de précision et de rappel sont très utilisés en recherche d'information textuelle (Coret A., Kremer P., Landi B., Schibler D., Schmitt L., Viscogliosi N. , 1997, Accès à l'information textuelle en français : le cycle exploratoire Amaryllis. Actes *1ères Journées Scientifiques et Techniques du réseau FRANCIL*, Avignon, France. 5-8).
Ils ont également été utilisés en étiquetage morphosyntaxique dans le cadre de l'action GRACE (Adda G., Mariani J., Paroubek P., Rajman M. et Lecomte J.,1999, L'action GRACE d'évaluation de l'assignation des parties du discours pour le français, *Langues*, 2(2), 119-129).
Le projet ARCADE a également défini des taux de précision et de rappel pour l'alignement des corpus multilingues **(**Véronis J., 1997, Une action d'évaluation des systèmes d'alignement de textes multilingues, Actes *1ères Journées Scientifiques et Techniques du réseau FRANCIL*, Avignon, France. 191-198 ; Véronis P., Langlais P., 2000, Evaluation of parallel text alignment systems : ARCADE. In Véronis J. (Ed.) *Parallel Text Processing*, Kluwer Academic, Dordrecht, Pays-Bas. 369-388.

13 Voir en particulier les taux d'erreur de mots (*WER* : *word error rate*) utilisés en reconnaissance de la parole. Ce taux est calculé à partir du nombre de suppressions, insertions ou substitutions détectées entre l'énoncé reconnu (hypothèse) et celui attendu (référence). Une approche similaire est souvent utilisée en compréhension de parole

14 King M. (1995) *op. cit*. (page 491).

15 Cette trop grande généralité ne concerne pas le niveau de granularité du composant évalué, mais l'ensemble des données de test. Ces évaluations très bien correspondre à une approche de type boîte transparente (*glass box*) où les performances de chaque module du système sont évaluées séparément. Compte tenu de la complexité des interactions entre ces différents composants, il est également recommandé de procéder en parallèle à une de type boîte noire (*black box*), c'est à dire concernant le système dans sa globalité. Voir par exemple : Polifroni J., Seneff S., Glass J., Hazen T.J. (1998) Evaluation methodology for a telephone-based conversational system. Actes *1st International Conference on Language Resources and Evaluation*, *LREC'98*, Grenade, Espagne, 43-49.

16 Caelen J., Zeiliger J., Bessac M., Siroux J., Perennou G. (1997) *op. cit*. (citation page 215).

Une autre insuffisance des campagnes d'évaluation de performances réside dans leur manque de généricité. Elles sont effet réalisées sur des jeux de test qui ne sont représentatifs que de la tâche considérée. Or, l'hypothèse d'indépendance du domaine postulée par James Allen et ses collègues ne peut selon moi s'appliquer qu'à des contextes applicatifs très proches[17]. C'est également la conclusion de Lynette Hirschman, qui a pris une part importante aux programmes d'évaluation MUC et ATIS de la (D)ARPA américaine[18] :

> *« By sticking to only one domain, ATIS failed to address the critical natural language portability problem »*

Il est pourtant essentiel que les recherches en communication homme-machine ne se cantonnent pas à l'étude de quelques domaines d'application. Compte tenu de l'importance de l'évaluation dans les pratiques de l'ingénierie des langues, il est important que cette recherche de généricité se reflète également dans nos paradigmes de test.

1.3. Conclusion : pour une évaluation discriminante à fort ancrage linguistique

Ce bref état de l'art suggère que nos pratiques d'évaluation ne sont pas totalement à la mesure des enjeux de l'ingénierie des langues. Il semble donc intéressant de développer des méthodes de test complémentaires aux évaluations de performances actuelles. Ces nouvelles méthodologies doivent présenter un pouvoir diagnostic équivalent à celui des analyses de fichiers de *log* tout en restant objectives et systématiques.

Cet objectif peut être atteint par une évaluation discriminante qui met en jeu des batteries de test dédiées à des situations ou à des phénomènes linguistiques bien délimités. Je propose de caractériser ces phénomènes par des analyses des usages langagiers afin de rapprocher évaluation objective et satisfaction des besoins des utilisateurs, mais aussi d'atteindre une plus grande généricité dans l'évaluation. C'est dans cette perspective que se situent mes travaux sur l'évaluation des systèmes de compréhension de la parole. Je vais détailler dans la suite de ce chapitre deux nouveaux paradigmes d'évaluation (DCR et DEFI) que j'ai proposés avec d'autres collègues, en montrant en quoi ces propositions répondent en partie aux insuffisances relevées ci-dessus.

2. EVALUER LA COMPRÉHENSION AUTOMATIQUE DE LA PAROLE

2.1. Compréhension automatique de la parole et dialogue homme - machine

Les serveurs vocaux interactifs sont des systèmes complexes composés de plusieurs composants interdépendants, parmi lesquels se trouve un module de compréhension de la parole. En première approximation, on peut estimer qu'ils reposent sur une architecture générique[19] semblable à celle donnée en figure 3.1.

---

[17] Selon Allen, la complexité du langage ou du contrôle du dialogue est indépendante de la tâche considérée (*domain independence hypothesis*) : Allen J., Byron D., Dzikovska M. (2000) An architecture for a generic dialogue shell. *Natural Language Engineering*. 6 (3-4). 213-228.

[18] Hirschman L. (1998) Language understanding evaluations : lessons learned from MUC and ATIS, Actes *2nd Conference on Language Resources and Evaluation*, *LREC'98*. Grenade, Espagne, 117-122 (citation p. 121).

[19] Le caractère sériel de cette architecture n'est pas aussi affirmé dans tous les systèmes. Par exemple, certains systèmes intègrent compréhension et de reconnaissance de parole : Seneff S., Mc Candless M., Zue V. (1995) Integrating natural language into the word graph search for simultaneous speech recognition and understanding. Actes *3rd European Conference on Speech Communication and Technology* , *Eurospeech'95*, Madrid, Espagne, 1781-1784 ; Young S., Ward W. (1993) Semantic and pragmatically based recognition of spontaneous speech. Actes *3rd European Conference on Speech Communication and Technology*, *Eurospeech'9*3, Berlin, Allemagne. 2244-2247.

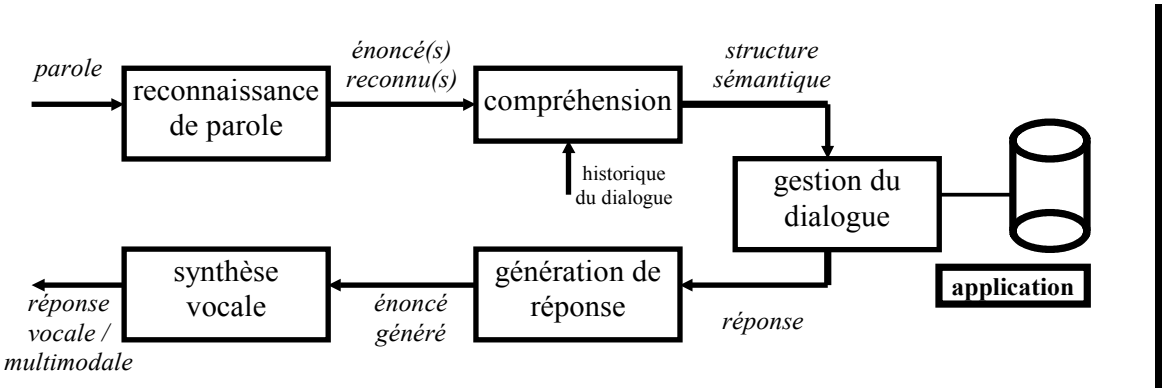

Figure 3.1 *— Architecture générale d'un système de dialogue oral homme - machine.*

Le message prononcé par l'utilisateur est tout d'abord traité par un module de reconnaissance de parole qui fournit en sortie une (ou plusieurs) solution(s) classée(s) par scores de vraisemblance décroissants. Ces solutions peuvent correspondre à des séquences de mots (énoncés reconnus) ou à un graphe de mots (on parle également de treillis de mots, ou *words lattice* en anglais).

Les énoncés reconnus sont alors traités par le module de compréhension afin que celui-ci construise leur représentation sémantique. Celle-ci est généralement obtenue en deux étapes. Tout d'abord, une étape de compréhension littérale consiste à caractériser et structurer l'information présente dans l'énoncé. Cette représentation sémantique est actualisée dans une seconde étape d'interprétation contextuelle qui va résoudre certaines ambiguïtés et co-références (anaphores, déictiques,…) en considérant l'historique du dialogue et de la tâche. L'objectif de cette seconde étape est également de caractériser les buts de l'utilisateur en identifiant les actes de dialogues présents dans l'énoncé.

La représentation sémantique obtenue est ensuite adressée au contrôleur de dialogue qui doit gérer :

- l'*interface avec l'application* — Dans le cas des systèmes d'information interactifs, le contrôleur de dialogue interroge la base de données à l'aide de requêtes SQL. Chaque interrogation fournit en retour un ensemble de réponses qui seront présentées à l'utilisateur.
- la *gestion du dialogue* — Le contrôleur de dialogue doit également gérer la construction interactive de la solution recherchée par l'utilisateur. Pour cela, il présente à l'utilisateur les réponses extraites de la base de données, lui demande des éclaircissements ou essaie de devancer ses requêtes. Il décide donc de la stratégie de dialogue à suivre compte tenu de la situation courante. A cette fin, il a en charge la maintenance des représentations qui rendent compte de l'état du dialogue et de la tâche. Aussi est-il fréquent qu'on lui confie également l'interprétation contextuelle des énoncés[20].

Le contrôleur de dialogue gère également la génération profonde des réponses à l'utilisateur. Ces sorties sont produites par l'intermédiaire d'une chaîne de génération qui varie suivant la modalité retenue. Dans le cas d'un serveur vocal interactif, on utilise une synthèse de parole à partir du texte[21] (*TTS : text-to-speech synthesis* en anglais).

2.2. Evaluation de la compréhension : l'influence du programme ATIS

La compréhension de la parole a déjà fait l'objet de campagnes d'évaluation d'envergure dans le cadre du programme *Speech and Natural Language* de la (D)ARPA américaine,[22]. Ce programme s'est intéressé à l'évaluation objective des systèmes de dialogue dans le cadre du renseignement aérien (ATIS). L'évaluation a porté sur les performances globales des systèmes (évaluation de type

[20] Rosset S. (2000) *op. cit*. La compréhension contextuelle est en effet profondément ancrée dans le dialogue

[21] Dutoit T. (1997) An introduction to text-to-speech synthesis. Kluwer Academic Publ., Dordrecht, Pays-Bas.

[22] Bates M., Boisen S., Makhoul J. (1992) Developing an evaluation methodology for spoken language systems. Actes *DARPA Speech and Natural Language Workshop*. 102-108 ; Hirschman L. *et al*. : MACDOW Group (1992) Multi-Site Data Collection for a spoken language Corpus. *Actes DARPA Speech and Natural Language Workshop*. 7-14.

boite noire) et sur celles des modules de reconnaissance et de compréhension[23]. L'évaluation de la compréhension n'a porté que sur l'interprétation littérale (hors contexte) des énoncés[24].

Les systèmes de compréhension de la parole peuvent reposer sur des approches relativement différentes. Afin de gérer cette diversité sans biais, l'évaluation ATIS a porté non pas sur les représentations sémantiques produites par les systèmes, mais sur les réponses obtenues après interrogation de la base de données. L'idée est que la compréhension correcte d'une requête conduira à une interrogation cohérente.

Tableau 3.2 *— Evaluation ATIS de la compréhension à travers les réponses globales du système. Exemple[25] de références correspondant à la requête : « Quels sont les tarifs sur les vols Paris /-Toulouse arrivant avant 15h ? »*

Référence Minimale		**Référence Maximale**		
code-tarif	no-vol	code-tarif	tarif	no-vol
Plein_Ciel	AF_2137	Plein_Ciel	95 Euros	AF_2137
Azur	AF_2137	Azur	125 Euros	AF_2137

D'une manière plus précise, chaque requête est associée à deux réponses canoniques appelées référence minimale et référence maximale (tableau 3.2). La référence minimale contient l'ensemble des informations attendues explicitement par l'utilisateur, tandis que la référence maximale contient des données supplémentaires jugées cohérentes avec la requête. La réponse du système est jugée correcte si elle contient la référence minimale et n'excède pas la référence maximale. On observe donc que l'évaluation ATIS se limite à la détection du sens strictement utile de l'énoncé.

Deux évaluations sont par ailleurs conduites simultanément afin d'étudier les interactions entre reconnaissance et compréhension de parole :

- Evaluation de la **compréhension langagière** (*natural language understanding*) — on considère en entrée la transcription de l'énoncé prononcé et en sortie la réponse de la base de données.
- Evaluation de la **compréhension de la parole** (*spoken language understanding*) — on considère en entrée le signal de parole et en sortie la réponse de la base de données.

Dans les deux cas, les résultats de l'évaluation reposent sur une métrique quantitative globale. On calcule un taux d'erreur sur l'ensemble du corpus de test. Il est estimé par le rapport entre le nombre de requêtes restées sans réponses ou ayant conduit à une réponse incorrecte et le nombre d'énoncés de test[26]. Nous sommes donc en présence d'une évaluation non discriminante des performances.

D'autres propositions ont été faites pour améliorer ce paradigme. En particulier, Samir Bennacef note qu'un test sur les réponses de la base de données est susceptible de biaiser l'évaluation de la compréhension[27]. Se basant sur son expérience des tests ATIS, il observe qu'une représentation sémantique erronée peut conduire à une réponse correcte de la base de données. C'est pourquoi il propose d'utiliser une représentation sémantique canonique comme référence de l'évaluation.

Cette solution nécessite toutefois la définition d'un système de représentation commun à tous les participants. Le portage vers cette représentation commune n'est pas anodin. On peut en particulier

---

23 Hirschman L. (1998) *op. cit.*

24 Les énoncés de tests recueillis avaient été divisés en trois catégories, :
- *énoncés de type A* — Enoncés dont la compréhension est indépendante du contexte dialogique.
- *énoncés de type D* — Enoncés dont l'interprétation dépend du contexte dialogique.
- *énoncés de type X* — Enoncés jugés hors du domaine de l'application.

L'évaluation n'a finalement porté que sur les énoncés de type A.

25 Exemple adapté de : Minker W. (1998). Evaluation methodologies for interactive speech systems. Actes *1st International Conference on Language Resource and Evaluation*, *LREC'98*, Grenade, Espagne, 199-206.

26 Une pondération d'un facteur deux est en fait effectuée entre les énoncés sans réponse et les énoncés incorrects.

27 Bennacef S. (1995) Modélisation du dialogue oral homme - machine : mise en œuvre dans une application de demande d'informations. Thèse de Doctorat, Université Paris XI, Orsay, France.

se demander si la représentation commune retenue est réellement neutre vis-à-vis des approches utilisées par chacun.

Quoi qu'il en soit, nous sommes toujours en présence d'une évaluation globale des performances. On retrouve donc avec les évaluations de type ATIS les limitations dont nous avons discuté plus haut. C'est pour répondre à ces insuffisances, sans perdre de vue la nécessité d'une évaluation objective, que nous avons proposé les paradigmes de test DCR et DEFI.

2.3. Evaluation DCR (Demande – Contrôle – Réponse)

Le paradigme DCR[28] a été initialement proposé par Jérôme Zeiliger (ICP), Jean Caelen (CLIPS-IMAG) et moi-même dans le cadre de l'action de recherche « Dialogue Oral » (ARC B2) de l'AUF (ex-AUPELF-UREF). Je me suis ensuite concentré sur les évolutions de la méthodologie au niveau de la compréhension de parole, tandis que Jean Caelen et Jacques Siroux (IRISA-CORDIAL) réfléchissaient à la généralisation du paradigme sur les niveaux dialogiques.

2.3.1. Objectifs — Le paradigme DCR répond à trois principes fondateurs :

- **Analyse discriminante** — L'évaluation DCR est de type diagnostic. Elle repose sur la définition de sous-sessions d'évaluation (séries de tests) qui se concentrent sur des phénomènes ou des procédés linguistiques précis[29]. J'ai rappelé (cf. chap. 2, § 1) l'intérêt des analyses d'usages pour la préparation des campagnes d'évaluation. Or, les études de corpus que j'ai réalisées montrent que la structure des procédés étudiés (dislocations, réparations, etc.) est relativement stable quel que soit le domaine d'application retenu (cf. chap. 2, § 3). Une évaluation s'appuyant sur de telles classes de phénomènes peut donc atteindre un certain degré de généricité[30].
- **Analyse objective** — L'évaluation DCR est par ailleurs objective. Elle repose sur des métriques quantitatives calculées sur chaque type de phénomènes. Le paradigme DCR se situe donc dans la classe des évaluations de type diagnostic définie par Hirschman et Thompson.
- **Evaluation sur les sorties de la compréhension** — L'évaluation DCR porte sur les sorties de la compréhension comme le recommandent les conclusions de Samir Bennacef[31]. Afin de ne pas recourir à l'adoption d'une représentation sémantique commune, l'évaluation ne repose pas sur une référence en sortie mais sur un énoncé de contrôle en entrée : c'est la représentation sémantique qu'en tire chaque système en interne qui tient lieu de référence à chacun.

A ma connaissance, ces principes ont été formulés pour la première fois dans le cadre du projet FRACAS (*FRAmework for Computational Semantics*) d'évaluation de la compréhension de textes écrits[32]. Cette méthodologie, qui fut reprise dans le cadre de l'ARC « Compréhension de texte »[33] de l'AUF, repose sur la définition de séries de tests spécifiques à des phénomènes linguistiques précis. Chaque test est constitué d'une déclaration D (énoncé à comprendre), d'un question fermée Q qui

[28] A l'origine, cette méthodologie avait reçu le nom de DQR (Donnée – Question – Réponse) : Zeiliger J., Caelen J., Antoine J.-Y. (1997) Vers une méthodologie d'évaluation qualitative des systèmes de compréhension et de dialogue oral homme - machine. Actes *1ères Journées Scientifiques et Techniques du réseau FRANCIL*, *JST-FRANCIL'97*, Avignon, France. 437-446 ; Texte repris dans : Chibout K., Mariani J., Masson N., Néel F. (Dir.) (2000) *Ressources et évaluations en ingénierie des langues. De Boeck Université*, Duculot, Bruxelles, Belgique. 437-461.

[29] Au début des années 1990, ces approches par suite de tests ont été envisagées dans de nombreux projets relevant généralement du TAL écrit. Outre le projet FRACAS décrit ci-après, on citera par exemple le projet TSNLP (*Test Suites for Natural Language Processing*) : Estival D *et al*. (1994) Survey of existing Test Suites, rapport de recherche du LRE 62-089 D-WP1, University of Essex, Royaume-Uni ; Lehmann, D. Estival, Oepen S. (1996). TSNLP : des jeux de phrases test pour l'évaluation d'applications dans le domaine du TAL, Actes *TALN 96*, Marseille, France. 97-103.

[30] Le croisement entre le diagnostic du comportement du système et une analyse des usages sur un nouveau domaine devrait fournir des indications fiables sur l'adaptation du système à ce nouveau contexte applicatif.

[31] Bennacef S. (1995) *op. cit*.

[32] FRACAS consortium. (1996). Using the framework. *Fracas project*. *LRE 62-051*, Deliverable D16 (chapitre 3).

[33] Sabatier P. (1997) Evaluer les systèmes de compréhension de textes, actes *1ères Journées Scientifiques et Techniques du réseau Francil*, *JST-FRANCIL'97*, Avignon, France, 223-226.

porte directement sur l'énoncé D et enfin d'une réponse attendue R (référence) à cette question Q. Voici un exemple d'évaluation de résolution d'anaphore proposé par FRACAS :

(D) *Peter is attending a meeting. He is to chair it.*

(Q) *Is Peter to chair a meeting ?*

(R) [Yes]

L'évaluation consiste à comparer la réponse fournie par le système à la question (Q) avec la référence (R). L'originalité de cette méthodologie réside dans l'introduction de la question (Q), qui déplace l'objet de l'évaluation. Il n'est plus nécessaire ici de comparer la structure sémantique construite à partir de (D) avec une référence prédéfinie. C'est au contraire la question (Q) qui impose au système une évaluation interne. En théorie, la comparaison avec la réponse (R) est neutre vis-à-vis des représentations, mais aussi des techniques utilisées par la compréhension.

La méthodologie DQR n'est toutefois pas directement transposable en communication orale homme-machine. Les systèmes de dialogue sont en effet conçus pour comprendre les requêtes de l'usager (demande D), et non pas pour s'interroger sur leur propre comportement (question Q) !

Aussi avons nous élaboré le paradigme d'évaluation DCR (Demande – Contrôle – Réponse) qui conserve les aspects intéressants de DQR sans nécessiter de telles capacités d'auto-analyse. C'est ce caractère original qui fait l'intérêt de cette méthodologie que je vais maintenant présenter en détail.

2.3.2. Le paradigme DCR : définition des jeux de tests

Tout comme DQR, le paradigme DCR repose sur la définition de batteries de tests. Chaque test DCR se compose de trois éléments :

- la **demande (D)** — Elle correspond à une requête de l'utilisateur sur laquelle va porter le test,
- le **contrôle (C)** — Il correspond à une reformulation ou simplification de l'énoncé (D) qui se focalise sur une information présente dans cette requête.
- la **référence (R)** — Elle est issue de la comparaison de (D) et (C). Elle est positive si les deux énoncés sont compatibles d'un point de vue sémantique et négative dans le cas contraire.

Lors de l'évaluation, le système doit donc interpréter les énoncés (D) et (C) pour tenter d'arriver au même jugement de compatibilité que la référence (R).

L'originalité du paradigme DCR repose dans la définition de l'énoncé de contrôle (C). A la différence de l'évaluation DQR, cet énoncé n'est pas une question réflexive qui interroge le système sur son propre comportement. Il s'agit au contraire d'une requête simplifiée qui pourrait avoir été prononcée par un utilisateur et qui vise à contrôler la bonne compréhension d'une information précise au sein de la demande (D). En conséquence :

- l'introduction de l'énoncé de contrôle (C) dans la procédure de test ne nécessite aucune modification du système,
- l'énoncé (C) doit être aussi simple que possible afin d'être compris sans faute par le système. Dans le cas contraire, la comparaison avec la référence (R) serait bien entendu biaisée.

A titre illustratif, reprenons l'exemple du tableau 3.2. Plusieurs tests DCR peuvent être définis, qui vérifient la compréhension des différents éléments de l'énoncé. On remarquera que ces tests peuvent être positifs ou négatifs, suivant la compatibilité des énoncés (D) et (C) :

(1) D *Quels sont les tarifs sur les vols Paris / Toulouse arrivant avant 15h ?*
C *Quelles sont les prestations ?*
R [Non]

(2) D *Quels sont les tarifs sur les vols Paris / Toulouse arrivant avant 15h ?*
C *Quels sont les tarifs ?*
R [Oui]

(3) D *Quels sont les tarifs sur les vols Paris / Toulouse arrivant avant 15h ?*
C *Quels sont les tarifs sur les Toulouse / Paris ?*
R [Non]

2.3.3. Le paradigme DCR : session d'évaluation

Chaque session d'évaluation DCR se déroule en trois étapes :

1) **Traitement séparé des énoncés (D) et (C)** — L'évaluation DCR se traduit par la compréhension parallèle d'énoncés indépendants. De ce point de vue, les approches ATIS et DCR sont équivalentes. DCR implique simplement un doublement des énoncés de test.

Tableau 3.3 *— Evaluation DCR : comparaison par unification des représentations sémantiques des énoncés (D) et (C). Cas d'une représentation sous forme de cadres sémantique (frame)* [34].

(D) : *Quels sont les tarifs sur les vols Paris / Toulouse arrivant au plus tard à 15h ?*	(C) *Quel sont les tarifs* ?	Unification (D),(C)
`type_requete : tarif`	`type_requete : tarif`	
`iti.depart : Paris`	`iti.depart : _`	
`iti.arrivee : Toulouse`	`iti.arrivee : _`	`[oui]`
`h.depart : _`	`h.depart : _`	
`h.arrivee : < 15.00`	`h.arrivee : _`	

2) **Comparaison des structures sémantiques de (D) et (C)** — Cette comparaison consiste à juger si les représentations produites par le système sont cohérentes[35]. Elle repose sur l'unification de ces représentations (tableau 3.3.). Il s'agit donc d'une évaluation interne qui est propre au système de représentation du système. Cette solution garantit la généricité de l'évaluation vis-à-vis des méthodes de compréhension et évite la définition, toujours difficile, d'une représentation sémantique commune à tous les participants.

3) **Comparaison avec la référence (R)** — Cette dernière étape décide de la validité des traitements effectués par le système. La compréhension est jugée correcte si le jugement de compatibilité issu de la comparaison précédente est identique au résultat attendu (R).

2.3.4. Le paradigme DCR : calcul et analyse des résultats

Les résultats de l'évaluation sont regroupés par séries de tests. Ils fournissent un taux d'erreur objectif pour chaque type de phénomène.

Plus précisément, chaque test est caractérisé par un ensemble de propriétés afin de permettre une analyse multicritère du comportement du système. En théorie, l'évaluation DCR peut ainsi fournir un diagnostic reposant sur un ensemble d'indices très discriminants[36]. Nous n'allons pas détailler ici l'ensemble des propriétés retenues[37]. Rappelons simplement les différents critères auxquelles elles correspondent :

---

[34] Pour un exemple, parmi bien d'autres, de systèmes de compréhension basé sur l'utilisation de cadres sémantiques : Oerder M. & Aust H. (1994) A real-time prototype of an automatic inquiry system, Actes *International Conference on Spoken Language Processing*, *ICSLP'94*, Yokohama, Japon, 703-706.

[35] Rappelons qu'un jugement négatif n'est pas synonyme de compréhension incorrecte, puisque que (D) et (C) peuvent être incompatibles (test négatif). La comparaison de ce jugement avec la (R) qui décidera de l'issue de l'évaluation.

[36] En pratique, la précision de l'évaluation dépend bien entendu de la taille et de la diversité de la banque de tests disponibles. Il s'agit d'une (fausse) limitation commune à l'ensemble des paradigmes diagnostiques (cf. § 2.3.6).

[37] Pour avoir plus de détails sur ce point, on consultera : Antoine J-Y., Caelen J. (1999) Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR (Demande, Contrôle, Résultat), *Langues*, 2(2). 130-139 ; Antoine J-Y., Siroux J., Caelen J., Villaneau J., Goulian J., Ahafhaf M. (2000) Obtaining predictive results with an objective evaluation of spoken dialogue systems : experiments with the DCR assessment paradigm, Actes *2nd Conference on Language Resources and Evaluation*, *LREC'2000*, Athènes, Grèce.

- *type d'énoncé* — Enoncé dont la compréhension est indépendante du contexte (type A dans la typologie ATIS), énoncé contextuel nécessitant soit la prise en compte de l'historique du dialogue, soit l'état courant de la tâche.
- *type sémantique du test* — Ce critère décrit la nature de la partie testée de l'énoncé (acte de dialogue, objet principal de la requête, modalité etc.),
- *complexité syntaxique de l'énoncé* — Ce critère décrit la présence de structures complexes telles que les subordonnées ou les coordinations,
- *parole spontanée* — Ce critère décrit la présence de procédés dus au caractère spontané de l'élocution (hésitations, répétitions, corrections, etc.)
- *référence* — Ce critère décrit la présence de co-références (anaphores, déictiques, etc.) ainsi que la nature de l'objet référencé (nombre, caractère défini ou non).
- *problèmes technologiques en entrée* — Ce critère décrit la présence d'erreurs de reconnaissance.

Chaque test est encodé sous un format structuré de type SGML (figure 3.2). Ce codage permet l'extraction automatique de sous-bases de tests pour des sessions d'évaluation spécifiques.

```
<test no="20_1" ctxt="HCTX" info="ACT" synt="SPL" tref="NON" nref="NON"
  oral="NON" >
<D> quel est le chemin pour se rendre à la Tour Eiffel </D>
<C> quel est le chemin </C>
<R>TRUE</R>
</test>
<test no="20_2" ctxt="HCTX" info="ACT" synt="SPL" tref="NON" nref="NON"
  oral="NON" >
<D> quel est le chemin pour se rendre à la Tour Eiffel </D>
<C> combien cela coûte-t-il </C>
<R>FALSE</R>
</test>
```

Figure 3.2 *— Exemples de tests DCR avec leur annotation*

2.3.5. Validation expérimentale de la méthodologie

La méthodologie DCR a été validée expérimentalement au cours d'une campagne de test qui n'a concerné qu'un seul système. Il s'agissait de LAMBDACOMP, le premier prototype du système de compréhension LOGUS développé au VALORIA (cf. chapitre 4 § 1.5.4). L'objectif de cette expérimentation était double. Il s'agissait tout d'abord de mettre en place une méthodologie de constitution systématique de tests DCR. Par ailleurs, il importait de vérifier que l'évaluation par comparaison demande (D) / contrôle (C) ne présentait aucun biais.

Cette expérimentation a porté sur 251 énoncés de test. Elle a été concluante d'un point de vue méthodologique et a dissipé tout doute sur l'existence de biais éventuels. En particulier, elle a permis de vérifier que la compréhension de l'énoncé de contrôle (C) ne faussait pas le processus d'évaluation. Sur les 251 tests, un seul énoncé de contrôle a donné lieu à une interprétation erronée. Ce cas isolé correspondait en fait à un test hors domaine, non identifié comme tel lors de la mise en place de l'expérimentation.

Tableau 3.4 *— Evaluation DCR du système LAMBDACOMP : taux d'erreurs de phrase suivant différents critères.*

Complexité syntaxique	Parole spontanée	Général
énoncés simples : 9,4 %	énoncés sans procédé de l'oral spontané : 4,4%	**9,6 %**
énoncés avec coordinations : 16,4 %	énoncés avec répétitions : 11,1 %	
subordonnées : non significatif	énoncés avec corrections : 66,6 %	

Compte tenu du nombre limité de tests, il ne nous a pas été possible de mener une analyse multicritères poussée sur les résultats de l'évaluation. Néanmoins, cette expérimentation a fourni des conclusions qui nous renseignent sur les capacités de diagnostic du paradigme DCR.

Ainsi, si le prototype LAMBDACOMP présente un taux d'erreur global de 9,6 %, la distribution des échecs suivant le type d'énoncé permet de mieux saisir les faiblesses du système (tableau 3.4). On constate par exemple un accroissement significatif des erreurs (16,4 %) dans le cas d'énoncés complexes comportant une coordination[38]. Par ailleurs, on remarque que les procédés relevant de l'oral spontané sont une source importante d'erreurs. Plus précisément, ce sont les corrections, plus que les répétitions qui posent problème à ce prototype.

Comme le montrent ces exemples, le paradigme DCR peut fournir des indications précises sur le fonctionnement des systèmes, alors qu'une évaluation globale de performances masque totalement ces renseignements. Ces informations ont un pouvoir prédictif au moins égal à celui des analyses de fichier de log. Les conclusions de cette campagne de test ont ainsi guidé de manière significative les évolutions du système LOGUS, qui répond désormais aux cas d'échecs les plus significatifs.

2.3.6. Limitations du paradigme DCR

L'expérimentation que je viens de présenter nous renseigne sur l'intérêt de l'évaluation DCR pour la conduite des recherches en dialogue homme-machine. A l'opposé, sa mise en œuvre relativement restreinte (251 tests) ne permet pas de juger complètement des limites de la méthodologie. Plutôt que de s'en remettre à un exercice d'auto-critique, il me semble intéressant de discuter dans ce mémoire des réactions d'autres chercheurs à nos propositions.

La méthodologie DCR a en effet rencontré un certain écho auprès de différents acteurs du domaine. Après avoir noté que Marcela Charfuelán et ses collègues de l'Université de Madrid relèvent essentiellement l'importance et la nécessité d'une évaluation discriminante telle que DCR[39], je vais m'intéresser aux problèmes soulevés par d'autres auteurs.

Généricité vis-à-vis des techniques d'analyse — Dans son mémoire de doctorat[40], Sophie Rosset (LIMSI) insiste également sur la contribution du paradigme DCR à la mise en œuvre d'une évaluation diagnostic. Elle note toutefois que cette méthodologie peut s'avérer dépendante du système. Cette limitation retient également l'attention de Laurence Devillers, Hélène Maynard et Patrick Paroubek, du même laboratoire[41].

Mon expérience m'inciterait à répondre que ce problème est réel mais limité. Plus précisément, ce manque de généricité provient de la constitution des énoncés de contrôle (C). Il est en effet nécessaire de considérer quelque peu le fonctionnement des systèmes évalués pour s'assurer que les énoncés de contrôle seront compris correctement. Dans le cas d'un système atypique (ou novateur !), il est effectivement possible que l'énoncé de contrôle pose problème. De même, le calcul de la référence par unification n'est pas aussi neutre qu'on pourrait le penser a priori. Du fait de la définition d'une représentation sémantique commune, ces limitations se retrouvent toutefois — avec certainement plus d'amplitude — dans les campagnes d'évaluation de performances.

---

[38] Une analyse plus fine des cas d'échec montre qu'il s'agit le plus souvent de requêtes multiples coordonnées. Ce type d'énoncé pose souvent problème aux systèmes de compréhension de parole : Minker W., Bennacef S. (1996) Compréhension et évaluation dans le domaine ATIS. Actes *XXI° Journées d'Etudes sur la Parole*, *JEP'1996*, Avignon, France. 415-419 (évaluation page 419).

[39] Charfuelán M., López C. E., Gil J. R., Rodríguez, Gómez L. H. (2000) A general evaluation framework to assess spoken language dialogues systems : experience with call center agent systems. Actes *TALN'2000*, Lausanne, Suisse.

[40] Cette critique, qui vise avant tout les niveaux d'évaluation du dialogue, peut s'étendre à la compréhension. On ne retiendra par contre pas la remarque selon laquelle DCR nécessite le développement d'un module capable d'analyse la « *question (Q)* ». Comme nous l'avons précisé, cette limitation du paradigme DQR proposé par FRACAS ne se retrouve plus dans DCR : Rosset S. (2000) Stratégies et gestionnaire de dialogue pour des systèmes d'interrogation de bases de données à reconnaissance vocale. Doctorat Université Paris XI, Orsay, France. publié comme rapport de recherche 2001-18 du LIMSI-CNRS, Orsay, France. septembre 2001.

[41] Devillers L., Maynard H., Paroubek P. (2002) Méthodologies d'évaluation du dialogue parlé : réflexions et expériences autour de la compréhension. *TAL*, vol. 43, n° 2, 155-184. (remarque p. 166).

Généricité vis-à-vis du domaine d'application — Une autre limitation vient du fait que les jeux de tests restent associés à une tâche donnée. On peut espérer que les conclusions de l'analyse discriminante (capacité à traiter les répétitions, par exemple) présentent un certain degré de généricité. Les analyses d'usages que j'ai présentées au chapitre précédent semblent conforter cet espoir. Il reste cependant difficile d'évaluer la part du spécifique et du générique dans ces résultats. Je pense toutefois que ce manque de généricité vis-à-vis de la tâche devrait être beaucoup moins sensible que dans le cas d'une évaluation de type ATIS.

Représentativité de l'évaluation — Dans un article ne portant pas sur la compréhension de parole, Salah Aït-Mokhtar et ses collègues (Xerox XRCE) notent qu'une évaluation fondée sur des séries de test ne permet pas de prédire le comportement des systèmes en situation réelle[42]. Cette affirmation, qui se fonde sur une expérimentation réalisée par Prasad et Sarkar, demande à être précisée[43]. Dans l'esprit des auteurs, « comportement réel du système » semble en effet signifier « performances globales ». De ce point de vue, la position du paradigme DCR est claire : il ne permet pas de relier évaluation diagnostic et évaluation des performances comme PARADISE le fait pour le diptyque évaluation objective / évaluation subjective. Ce n'est pas son objectif. A l'opposé, la connaissance du comportement des systèmes sur un phénomène précis est intéressante dans toute situation réelle où il se retrouve…

Aït-Mokhtar et ses collègues avancent un autre argument sur le manque de représentativité des méthodologies d'évaluation diagnostic. Ils relèvent que[44] :

> *« chaque phrase de test illustre généralement un phénomène spécifique, alors que la combinaison de plusieurs phénomènes complexes est courante dans les textes réels »*

Il s'agit là d'un problème réel, qui est fréquemment ignoré des évaluations par série de tests. Le paradigme DCR tente de répondre à cette limitation en autorisant la présence de plusieurs phénomènes dans un test particulier. Plusieurs propriétés (cf. § 2.3.4) peuvent ainsi être associés à un test donné, l'analyse multicritère des résultats devant permettre de dissocier ou au contraire de corréler les causes d'erreurs du système.

Compréhension contextuelle — A la suite des remarques et propositions de Hélène Maynard et Laurence Devillers (LIMSI) sur le paradigme PEACE[45], je ferai remarquer que le paradigme DCR se prête difficilement à une évaluation de la compréhension contextuelle. Dans le cas d'énoncés dont l'interprétation requiert la prise en compte de l'historique du dialogue, nous avons proposé de construire un énoncé complexe (D) regroupant l'ensemble des tours de paroles précédents. Cette solution s'avère en pratique problématique :

- soit l'énoncé (D) ne contient que les tours de parole correspondant à l'usager, et une partie de l'information est manquante,
- soit l'énoncé (D) intègre les réponses du système, mais celui-ci n'est alors pas conçu pour analyser ses propres productions.

Cette limitation n'est cependant pas consubstantielle à l'évaluation DCR et peut être contournée. L'astucieuse technique de paraphrasage du dialogue qu'ont proposé Hélène Bonneau-Maynard et Laurence Devillers pour PEACE peut en effet être utilisée par la méthodologie DCR.

Coût de l'évaluation diagnostique — Au final, le seul problème réellement incontournable

---

42 Aït-Mokhtar S., Hagège C., Sándor Á. (2003) Problèmes d'intersubjectivité dans l'évaluation des analyseurs syntaxiques. Actes *Atelier TALN'2003 sur l'évaluation des analyseurs syntaxiques*. Batz-sur-Mer, France, Vol. 2, 57-66.

43 Prasad R., Sarkar A. (2000) Comparing test-suite based evaluation and corpus-based evaluation of a wide-coverage grammar for English. Actes *2nd International Conference on Language Resource and Evaluation, LREC'2000 Workshop on using evaluation within HLT programs : results and trands*, Athènes, Grèce, 7-12

44 Aït-Mokhtar S. et al. (2003). *op. cit*, p. 58.

45 Devillers L., Maynard H., Paroubek P. (2002) *op. cit.* ; Maynard H., Devillers L. (2000) A framework for evaluating contextual understanding. Actes *6th International Conference on Spoken Language Processing, ICSLP'2000*. Pékin, Chine.

rencontré par le paradigme DCR semble être son coût de développement. La constitution systématique de séries de tests DCR demande en effet un effort important qui est d'autant plus élevé que les critères discriminants retenus sont nombreux. Cette difficulté a d'ailleurs certainement empêché la généralisation des autres paradigmes d'évaluation diagnostic (*TSNLP*, *FRACaS, etc…*). Afin de répondre à ce problème, Zakaria Kurdi et Mohamed Ahafhaf (CLIPS-IMAG et Université d'Odense) ont proposé d'étendre la méthodologie DCR par une génération automatique des énoncés de contrôle[46]. Intéressante au premier abord, cette solution ne me semble pas être à la mesure du problème. D'une part, la construction d'une « grammaire des erreurs » rendant compte des usages langagiers observés est certainement aussi lourde que la création d'une batterie de tests. D'autre part, si un système était capable de générer l'ensemble des usages langagiers que l'on désire tester, il résoudrait définitivement le problème de la compréhension sur la tâche étudiée !

2.3.7. Bilan : apport du paradigme DCR aux recherches en dialogue oral

Ce problème de coût constitue selon moi un faux problème. Je pense que la mise en œuvre de paradigmes lourds d'évaluation est un passage obligé si on veut disposer d'indications prédictives dans un cadre compétitif. Après des années de pratiques perfectibles en informatique, le génie logiciel a montré que l'effort que l'on doit accorder à la validation doit être aussi important que celui nécessité par le développement proprement dit. Au moment où se pose la question des « *best practises* » en ingénierie des langues[47], il est temps de faire nôtres ces recommandations.

L'AUPELF-UREF ayant décidé d'interrompre son soutien à la recherche francophone en ingénierie des langues, nous n'avons pas eu réellement l'opportunité d'ouvrir cette voie avec le paradigme DCR. Il est toutefois clair que la méthodologie DCR a eu un impact déterminant sur les recherches du domaine. Comme nous venons de le voir, plusieurs auteurs ont situé leurs propositions par rapport à cette méthodologie. De même, il est significatif que les dernières recherches francophones en matière d'évaluation de la compréhension ont toutes reposé sur une approche discriminante.

Je vais présenter ces travaux récents, en commençant par une campagne d'évaluation que j'ai animée pendant deux ans. Ce programme reposait sur une nouvelle méthodologie d'évaluation qui se voulait plus légère que DCR tout en reprenant ses principes fondateurs.

2.4. Campagne d'évaluation par défi

Cette méthodologie, appelée « évaluation par défi », a été développée à partir de septembre 2000 dans le cadre du groupe de travail « Compréhension de parole » (GT 5.5.) du GDR-I3 du CNRS. Je dirige depuis 1998 ce groupe qui réunit actuellement les laboratoires CLIPS-IMAG, IRIT, LIMSI, LORIA et VALORIA autour de la problématique de la compréhension de parole en dialogue homme - machine ou en recherche d'information (systèmes interactifs de type question / réponse).

2.4.1. Objectifs et principes

La méthodologie d'évaluation par défi que j'ai proposée et qui a ensuite été affinée par l'ensemble des participants reprend les objectifs principaux de DCR. Elle permet une confrontation objective du comportement des systèmes tout en étant plus centrée sur une capitalisation des connaissances que sur une compétition entre participants. Sa spécificité est ainsi de porter sur des jeux de tests différents (mais comparables) pour chaque système. Cette caractéristique interdit toute comparaison directe des performances. Elle permet par contre une certaine légèreté de mise en œuvre, puisqu'elle ne nécessite pas la définition d'une tâche ou même d'un système de représentation communs.

En pratique, l'évaluation par défi suit le processus suivant :

Jeux de tests spécifiques pour chaque système — Ces tests spécifiques sont élaborés à partir d'énoncés *initiaux* qui sont fournis par le concepteur du système et jugés représentatifs de la tâche.

[46] Kurdi M. Z., Ahafhaf M. (2002) Toward an objective and generic method for spoken language understanding systems evaluation : an extension of the DCR method. Actes *3rd International Conference on Language Resources and Evaluation*, *LREC'2002*. Las Palmas de Gran Canaria, Espagne. 545-550.

[47] van Kuppevelt, Heid U., Kamp H. (2000) Best practice in spoken language dialogue systems engineering. *Natural Language Engineering*, 6 (3-4). 205-212.

Défi — Les jeux de tests sont constitués d'*énoncés dérivés* construits à partir des énoncés initiaux. Ceux-ci sont élaborés par l'ensemble des participants, à l'exception du concepteur du système. Les énoncés dérivés sont une réécriture ou une complexification des énoncés initiaux. Ils sont supposés poser problème au système, d'où la notion de *défi*. Ainsi, l'énoncé dérivé (D) ci-dessous a été produit à partir de l'énoncé initial (I) :

(I) *non le matin à 6 heures environ*

(D) *non c'est le matin je euh à six heures environ que je voudrais partir*

Tout énoncé dérivé est validé par le concepteur du système qui vérifie s'il n'est pas hors périmètre. La dérivation est aussi systématique que possible. C'est-à-dire qu'on dérive un ensemble de séries de tests suivant une typologie préalablement définie de phénomènes (cf. infra). On retrouve ici le caractère discriminant de la méthodologie DCR.

Evaluation — Chaque système est évalué sur le jeu d'énoncés dérivés qui lui est propre. C'est le concepteur du système qui juge si les réponses du système sont correctes. Les résultats sont synthétisés par des mesures objectives (taux d'erreurs, distribution des erreurs) calculées sur chaque série de test. On retrouve le principe d'analyse diagnostic développé par la méthodologie DCR.

Bilan — Une journée d'étude est consacrée à l'analyse des résultats en fin de campagne. Cet atelier permet de comparer le comportement des systèmes au regard des techniques qu'ils utilisent.

2.4.2. Campagne d'évaluation par défi du GDR-I3 : typologie de phénomènes

Cette méthodologie a été validée en 2001 par une campagne d'évaluation grandeur nature qui a réunit les participants du groupe de travail « Compréhension de parole » à l'exception du LORIA.

Tableau 3.4 *— Systèmes de compréhension de parole participant à la première campagne d'évaluation par défi du GDR-I3 du CNRS.*

Système	Concepteur	Laboratoire	Domaine d'application
OASIS	M. Z. Kurdi[48]	CLIPS-IMAG	réservation hôtelière
CACAO	C. Bousquet-Vernhettes[49]	IRIT	renseignement horaires de train
ARISE	S. Rosset[50]	LIMSI	renseignement ferroviaire
ROMUS	J. Goulian[51]	VALORIA	renseignement touristique
LOGUS	J. Villaneau[52]	VALORIA	renseignement touristique

Chaque système a été évalué sur un nombre significatif de 1200 énoncés de test. Ces tests ont été produits à partir de vingt énoncés initiaux jugés représentatifs de la tâche retenue pour le système considéré (tableau 3.4). Chaque participant a alors produit 15 énoncés dérivés par énoncé initial.

[48] Kurdi M.Z. (2001) A spoken language understanding approach which combines the parsing robustness with the interpretation deepness, Actes *International Conference on Artificial Intelligence, ICAI'01*, Las Vegas, Etats-Unis.

[49] Bousquet-Vernhettes C. (2002) Compréhension robuste de la parole spontanée dans le dialogue oral homme - machine : décodage conceptuel stochastique. Thèse Université Paul Sabatier, Toulouse, France, septembre 2002.

[50] Rosset S., Lamel L. (2001) Gestionnaire de dialogue pour un système de dialogue à reconnaissance vocale. Actes *TALN'01*, Tours, France. 385-390.

[51] Goulian J., Antoine J.-Y. (2001) Compréhension automatique de la parole combinant syntaxe locale et sémantique globale pour une CHM portant sur des tâches relativement complexes, Actes *TALN'2001,* Tours, France. 203-212 ; Goulian J. (2002) Stratégie d'analyse détaillée pour la compréhension automatique robuste de la parole. Thèse Université de Bretagne Sud, Vannes, France. Décembre 2002.

[52] Villaneau J., Antoine J.-Y., Ridoux O. (2001) Combining syntax et pragmatic knowledge for the understanding of spontaneous spoken utterances. Actes *4th International Conference on the Logical Aspects of Computational Linguistics*, *LACL'01,* Le Croisic, France. In *LNAI 2099*, Springer Verlag, 279-295.

Tableau 3.5 — *Synthèse du comportement de chaque système observé par évaluation par défi (☺ : gestion satisfaisante des problèmes ; ✯ : gestion satisfaisante mais perfectible sur une sous-catégorie de problèmes ; ☹ : gestion des problèmes perfectible).*

Système	CLIPS (Oasis)	IRIT (Cacao)	LIMSI (Arise)	VALORIA (Romus)	VALORIA (Logus)
erreurs de RAP	☹	☺	☺	☹	☹
complexité	☹ (objets complexes)	☹ (requêtes multiples)	☹ (objets complexes)	✯ (portée négations)	✯ (requête+info)
oral spontané	☺	☺ (*sauf* incise dans segment)	☹ (incise hors domaine)	✯ (incise dans segment)	✯ (faux départs)
dislocations	☺	☹	☹	☺	☺
pbs de couverture	☹	☹ (mots utiles HV)			

Une réunion de synthèse, organisée à Toulouse en octobre 2001, nous a permis de faire le bilan de cette campagne de test. Les tableaux 3.5 et 3.6 synthétisent les résultats et problèmes rencontrés par chaque système suivant la typologie que nous avions définie[53]. La caractérisation des situations problématiques est restée à la discrétion de chaque concepteur de système. Ce tableau présente donc surtout les pistes d'amélioration qui paraissent prioritaires pour chaque système.

Etant donné que ce chapitre porte sur les pratiques d'évaluation, je ne m'attarderai pas sur l'analyse du comportement de chaque système[54]. Remarquons cependant que ce tableau de synthèse confirme le pouvoir prédictif de la méthodologie, puisqu'il permet une caractérisation détaillée des forces et faiblesses de chaque système.

Tableau 3.6 — *Campagne d'évaluation par défi du GDR-I3 du CNRS : distribution des erreurs rencontrés par chaque système.*

Systèmes	CLIPS (Oasis)	IRIT (Cacao)	LIMSI (Arise)	VALORIA (Romus)	VALORIA (Logus)
Erreurs de reconnaissance	7,0 %	0%	0 %	20%	2%
Enoncés complexes	12,5 %	6.5%	0 %	6%	8%
Procédés de l'oral spontané	9,0 %	6%	18,2 %	17%	32%
Dislocations	2,3 %	14.9%	9,0 %	6%	3%
Problèmes de couverture	69,2 %	72.6%	36,0 %	32%	35%
Autres (causes combinées)	0 %	0 %	36,8 %	19%	20%

On constate que ces points forts reflètent logiquement les motivations scientifiques des concepteurs des systèmes. Par exemple, le VALORIA cherche à développer des techniques d'analyse à fort ancrage linguistique. Il est donc encourageant que ses systèmes LOGUS et ROMUS soient capables

[53] Certains systèmes n'ont pas été évalués sur toutes les classes de problèmes, soit parce que les phénomènes considérés sortaient du périmètre de la tâche, soit parce qu'un diagnostic clair n'a pu être tiré.

[54] Pour une présentation plus détaillée, voir : Antoine J-Y. *et al*. (2001) Synthèse de la réunion d'analyse des résultats de la campagne d'évaluation par défi. Disponible sur la Toile: www.univ-ubs.fr/valoria/antoine/gdri3/Oct01.html ; Antoine J.-Y., Bousquet-Vernhettes C., Goulian J., Kurdi M. Z., Rosset S., Vigouroux N., Villaneau J. (2002) Predictive and objective evaluation of speech understanding: the "challenge" evaluation campaign of the I3 speech workgroup of the French CNRS. Actes *3rd Conference on Language Resources and Evaluation*, *LREC'2002*, Las Palmas de Gran Canaria, Espagne.
Pour une analyse complète du comportement des systèmes CACAO (IRIT), OASIS (CLIPS) et ROMUS, on se réfèrera respectivement à : Bousquet-Vernhettes C. (2002) *op. cit*. (pages 148-154) ; Kurdi M.Z. (2001) *op. cit*. ; Goulian J. (2002) *op. cit*. (pages 124-129).

de traiter des énoncés structurellement complexes. A l'opposé, le VALORIA a défié les autres participants avec des dérivations complexes qui sont marginales sur les domaines étudiés par l'IRIT et le LIMSI. Leurs systèmes se sont pourtant comportés correctement sur ces situations pour lesquelles ils n'avaient pas été conçus. Il s'agit d'une découverte prometteuse pour ces participants, qui s'accordent par ailleurs sur l'intérêt des phénomènes concernés.

L'évaluation par défi permet donc de tester un système sur des situations qui ne relèvent pas des préoccupations immédiates de ses concepteurs, mais qui pourraient avoir une importance croissante à l'avenir. Cette observation confirme le caractère prédictif de la méthodologie.

2.4.3. Conclusion : de l'importance d'une comparaison dépassionnée entre systèmes

Les enseignements que notre groupe de travail a tiré de cette première campagne d'évaluation par défi confirment mes propos sur l'intérêt d'une évaluation diagnostic. Comme pour la méthodologie d'évaluation DCR, cette campagne de test a en effet fourni des indications sur le comportement des systèmes bien plus précieuses que ce qu'aurait pu donner une évaluation globale de type ATIS. Par ailleurs, on observe que les problèmes retenus dans notre typologie définissent des réalités technologiques ou linguistiques transversales à tout domaine applicatif. L'évaluation par défi présente donc une certaine garantie de généricité.

Cette généricité est favorisée par l'absence de compétition entre les participants de l'évaluation par défi. Ici, les concepteurs n'ont guère intérêt à rechercher une spécialisation à outrance du système sur la tâche de test, au détriment d'évolutions plus intéressantes à long terme. Comme le rappellent Hirschman et Thompson, cet effet indésirable est au contraire consubstantiel aux évaluations de performances[55]:

> *« We have yet to develop good methods of evaluating* ***understanding*** *independent of* ***doing the right thing*** *in the context of a specific application* […] *performance evaluation may lead to risk-avoidance strategies where getting a good score becomes more important than doing good research »* (souligné par les auteurs)

L'évaluation par défi ne prétend pas résoudre cette contradiction. Remarquons simplement qu'en dépassionnant la confrontation entre systèmes, elle y est moins sujette.

En conclusion, cette première campagne d'évaluation par défi a montré le caractère prometteur de ce paradigme de test discriminant. Un des résultats les plus significatifs de ce travail est d'ailleurs d'avoir reçu l'adhésion de l'ensemble des chercheurs français en compréhension de parole autour de cette méthodologie novatrice.

Il n'en reste pas moins que les enseignements de cette campagne ont été limités du fait du manque de systématicité qui a présidé à la constitution des tests dérivés. Afin de répondre à cet impératif, notre groupe travaille désormais à une définition plus précise de la typologie de problèmes rencontrés par la compréhension de parole. Cette réflexion, qui regroupe tous les acteurs francophones en compréhension de la parole, devrait conduire à la définition d'un agenda des recherches futures du domaine. Elle sera d'ailleurs utilisée dans le cadre de la future campagne d'évaluation MEDIA à laquelle participera le VALORIA durant les années 2003-2004.

3. EPILOGUE : CAMPAGNE D'EVALUATION MEDIA

3.1. Le paradigme PEACE : évaluation de la compréhension contextuelle

La campagne MEDIA d'évaluation de la compréhension de la parole est organisée dans le cadre de l'appel d'offre TECHNOLANGUE du Ministère de la Recherche (programme EVALDA). Elle repose sur la méthodologie d'évaluation PEACE définie par Hélène Boneau-Maynard et Laurence Devillers[56]. Je tiens à présenter cette nouvelle méthodologie car elle témoigne de l'influence qu'ont

[55] Hirschman L., Thompson H.S. (1995) *op. cit.*

[56] PEACE = Paradigme d'Evaluation Automatique de la Compréhension hors et En-contexte : Devillers L., Maynard H., Paroubek P. (2002) *op. cit.* ; Bonneau-Maynard H., Devillers L. (2000) *op. cit*.

eu les propositions d'évaluation DCR et DEFI sur la communauté scientifique. Comme le rappelle d'emblée Laurence Devillers dans la présentation scientifique du projet MEDIA[57]:

> *«* [PEACE] *est fondé sur la constitution de batteries de tests reproductibles issues de dialogues réels . Ce paradigme suit le même courant d'idée que les évaluations DQR et DEFI basées sur des batteries de tests »*

Je ne peux que me réjouir de cette convergence entre les positions défendues dans ce mémoire et celles de chercheurs jusqu'ici coutumiers des évaluations globales de performances. La campagne MEDIA marquera ainsi une évolution significative par rapport aux évaluations de type ATIS du fait de son caractère diagnostic affirmé[58].

Un des intérêts de la campagne d'évaluation MEDIA est de porter sur l'interprétation en contexte et non plus seulement sur la compréhension littérale. Cette question, rarement abordée jusqu'à présent, devrait permettre une meilleure compréhension des interdépendances qui existent entre interprétation et contrôle du dialogue. Nous avons vu précédemment que la méthodologie DCR s'appliquait difficilement à la compréhension contextuelle. En reprenant l'idée d'un énoncé regroupant l'historique du dialogue déjà réalisé, les promoteurs du paradigme PEACE proposent une solution originale qui devrait limiter les biais méthodologiques rencontrés avec DCR.

Le principe est de fournir au système la succession des données à prendre en compte sous la forme d'une paraphrase du contexte. Cette paraphrase doit correspondre à un énoncé susceptible d'avoir été prononcé dans une situation normale. Les auteurs proposent de générer cet énoncé à partir d'une représentation sémantique qui rend compte de l'évolution du dialogue. Les expérimentations déjà réalisées semblent démontrer que cette génération n'introduit pas de biais dans l'évaluation[59].

3.2. Conclusion

En conclusion, la généralisation des campagnes d'évaluation objective a constitué un apport essentiel au développement de l'ingénierie des langues. Au cours de ce chapitre, j'ai cependant montré que les paradigmes de tests purement technologiques (évaluation des performances) souffrent d'un manque de pouvoir diagnostic. Il est apparu au cours de la discussion qu'on ne peut atteindre une évaluation prédictive que par la mise en place d'une évaluation discriminante du comportement du système sur les usages langagiers réels des utilisateurs.

Ce retour au linguistique a été initié dans le cadre de l'ARC « Dialogue Oral » de l'AUPELF-UREF avec la proposition de l'évaluation DCR. Il est désormais consacré par le groupe de travail du GDR-I3 que je dirige (évaluation par défi), mais aussi par la prochaine campagne d'évaluation MEDIA. Tous deux réunissent désormais l'ensemble des acteurs français de la compréhension de parole autour de propositions que j'ai fortement défendu au cours des cinq dernières années. Bien souvent, les moyens m'ont manqué pour pousser plus en avant ces propositions. C'est donc avec un grand plaisir que j'observe qu'elles ont néanmoins retenu l'attention et l'adhésion de la communauté scientifique.

Comme nous le voyons, les réflexions sous-jacentes à l'évaluation des systèmes d'ingénierie des langues révèlent des préoccupations épistémologiques beaucoup plus profondes que la simple validation des systèmes. C'est là tout l'intérêt d'une réflexion sur l'évaluation, qui ne doit pas être limitée à une compétition ou à une vérification d'hypothèses, mais être considérée comme une activité fondatrice de notre pratique scientifique. Une maxime pourrait résumer cette réalité : dis moi comment tu évalues et je te dirai ce qu'est ton système…

---

[57] Document non publié.

[58] L'objectif du projet MEDIA est en effet de fournir, à côté d'une mesure globale de performances, des indices discriminants sur le comportement sur des problèmes particuliers. Le paradigme PEACE repose par contre sur la définition — classique — d'une application et d'une représentation sémantique commune à tous les systèmes.

[59] Devillers L., Maynard H., Paroubek P. (2002) *op. cit*. Il est cependant clair qu'une reformulation du contexte n'est pas identique à sa perception directe sous forme de tours de parole successifs. Cette méthodologie ne fait donc qu'approcher l'évaluation de la compréhension contextuelle. Elle n'en constitue pas moins une avancée réelle dans cette direction.

C'est précisément ce que nous allons étudier dans le prochain chapitre, consacré aux réalisations du groupe CORAIL. On y observera l'influence des analyses d'usages que j'ai présentées sur la conception de nos systèmes.

4. Les réalisations du VALORIA : segments noyaux et TAL robuste

Comprendre c'est compliquer.
C'est enrichir en profondeur.

Lucien Febvre, *Combats pour l'histoire*.

Au cours de ce dernier chapitre, je vais présenter les systèmes réalisés sous ma direction au sein du groupe CORAIL du VALORIA. J'insisterai à cette occasion sur les principes fondateurs qui leur sont sous-jacents. Mon propos est en effet de montrer la cohérence scientifique qui existe entre les études amonts et avals que j'ai présentées précédemment et la conception de nos systèmes. Ces derniers font ou ont fait l'objet des doctorats de Jérôme Goulian, Jeanne Villaneau et Igor Schadle. Je ne peux qu'encourager le lecteur intéressé à consulter les mémoires de thèse de ces étudiants.

Je reviendrai dans un premier temps sur nos recherches en compréhension automatique de la parole. Dans la continuité de mon doctorat, ces travaux sont orientés vers la mise en œuvre d'une interprétation fine mais robuste des énoncés oraux spontanés. Je discuterai de l'intérêt d'une analyse à fort ancrage linguistique tout en montrant qu'elle constitue une démarche technologiquement viable. Nous retrouverons cette double préoccupation dans nos travaux consacrés aux systèmes d'aide aux handicapés. Cette partie sera l'occasion de constater que le TAL stochastique n'est pas nécessairement incompatible avec la recherche d'une modélisation linguistique fine.

1. COMPREHENSION DE LA PAROLE : DU TAL ROBUSTE AU TALP

La compréhension de parole est, avec la dictée vocale, le champ d'investigation du dialogue homme - machine où les approches purement ingénieriques sont les plus développées. Sans chercher à masquer les réussites de ces méthodes, je vais discuter ici de leurs limitations et présenter des solutions alternatives pour y remédier.

1.1. Compréhension de parole et dialogue finalisé : difficultés

J'ai présenté dans le chapitre précédent (cf. chap. 3 § 2.1) le rôle de la compréhension de parole en dialogue homme-machine. Par rapport à la compréhension de textes ou à l'analyse syntaxique du langage écrit, cette problématique induit des problèmes spécifiques de deux natures.

Tableau 4.1 *— Campagnes ARPA d'évaluation de la reconnaissance de la parole : performances obtenues par les meilleurs systèmes*[1]

Campagne	Elocution	Application	Taux d'erreur de mots
ATIS	Parole spontanée	Finalisée (2000 mots)	3%
Business news	Textes lus (*WSJ*)	Non finalisée (> 20 000 mots)	7,2 %

Gestion des erreurs de la reconnaissance automatique de la parole — Tout d'abord, le module de compréhension intervient après[2] l'étape de reconnaissance automatique de la parole. Il travaille donc sur des séquences de mots perturbés par de nombreuses erreurs de reconnaissance. La reconnaissance automatique de la parole a connu des avancées significatives au cours de la dernière

[1] Pallet D., Fiscus J., Fisher W., Garofolo J., Lund B., Prysboski M. (1994) 1993 benchmark tests for the ARPA spoken language program. Actes *1994 ARPA Human Language Technology workshop*. Morgan Kaufman, Princeton, NJ. 49-74.

[2] Ou plus rarement simultanément, comme par exemple dans les travaux du MIT ou de CMU : Seneff S., Mc Candless M., Zue V. (1995) Integrating natural language into the word graph search for simultaneous speech recognition and understanding, actes *4th European Conference on Speech Communication and Technology* , *Eurospeech'95*, Madrid, Espagne, 1781-1784 ; Issar S., Ward W. (1993) CMU's robust spoken language understanding system. Actes *3rd European Conference on Speech Communication and Technology*, *Eurospeech'9*3, Berlin, Allemagne. 2147-2151 ; Young S., Ward W. (1993) Semantic and pragmatically based recognition of spontaneous speech. Actes *3rd European Conference on Speech Communication and Technology*, *Eurospeech'9*3, Berlin, Allemagne. 2244-2247.

décennie. Comme le montre le tableau 4.1, les meilleurs systèmes présentent désormais des performances assez remarquables dans le cas d'applications très finalisées (ATIS), ou lorsque la reconnaissance porte sur de la parole lue (*Business News*). Ces progrès ne sont malheureusement pas au rendez-vous dans les situations plus difficiles. Les taux d'erreurs atteignent ainsi rapidement les 10 % et plus dans le cas de la transcription de journaux radiodiffusés (*Broadcast News*[3]). Ils grimpent au dessus des 25% dans le cas de conversations téléphoniques réellement spontanées (corpus *Switchboard*[4]) !

Pour donner une mesure de ce problème en dialogue homme-machine, Caroline Bousquet a défini la notion d'interprétabilité d'un énoncé transcrit . Un énoncé est dit interprétable si les erreurs de reconnaissance n'empêchent pas la compréhension, par un observateur humain, du sens de l'énoncé originel[5]. Elle montre sur les corpus du programme ARISE (renseignements d'horaires de trains) qu'un taux d'erreur de mots de 40,6 % conduit à une proportion d'énoncés non interprétables de 39,8 %. On comprendra que cette perte d'information, non détectable a priori, rend particulièrement difficile la tâche de la compréhension de parole et du contrôleur de dialogue.

Modélisation des procédés de l'oral spontané — D'autre part, le caractère spontané du langage parlé se traduit par la présence de nombreuses constructions qui cassent la structure des énoncés. Au cours de chapitre 2 (§ 3.3), nous nous sommes spécifiquement intéressés aux cas des réparations. Il convient d'y ajouter les hésitations, les incises ou encore les inachèvements.

Ces procédés gênent fortement la compréhension de la parole spontanée. Comme nous l'avons vu au chapitre 2 (§ 3.3.), il est possible de détecter certaines réparations à l'aide de techniques de pattern-matching. Si elles apportent une aide non négligeable au traitement du langage parlé, ces méthodes de détection superficielles ne peuvent rendre compte de l'ensemble des procédés de l'oral spontané.

C'est pour répondre à ces difficultés « résiduelles » — en fait les plus nombreuses ! — que les chercheurs en dialogue oral homme-machine ont développé des méthodes spécifiques à leur problématique. Une des voies les plus explorées consiste à profiter du caractère finalisé du dialogue pour développer des approches orientées par la tâche. Cette démarche a donné naissance aux techniques sélectives de compréhension de la parole.

1.2. Réussites de la compréhension sélective de la parole

Les techniques sélectives de compréhension de parole répondent à une logique ingénierique qui tente de faire face au difficile problème de robustesse que pose le traitement de la parole spontanée. Ces travaux reposent sur une idée simple : puisque certaines parties de l'énoncé résistent à l'analyse automatique, ignorons-les si c'est possible. La compréhension se limite alors à l'identification de quelques îlots clés qui représentent des unités de sens (les *segments conceptuels*) pertinentes dans l'univers de la tâche. Ces traitements sélectifs restreignent la compréhension à un sens qualifié

[3] A titre d'exemple, le taux d'erreur de mots du système *Broadcast News* du LIMSI est proche 14% : Lefèvre F., Gauvain J.-L., Lamel L. (2002) Développement d'une technique générique pour la reconnaissance de la parole indépendante de la tâche. Actes *XXIV° Journées d'Etudes sur la Parole*, *JEP'2002*, Nancy, France, 221-224.

[4] Cohen J., Gish H. et Flanagan J. (1994) Switchboard, the second year. Actes *CAIP summer workshop in speech recognition: frontiers in Speech Processing II*.

[5] Plus précisément : « *une solution donnée par un module de reconnaissance de la parole est dite interprétable si et seulement si la représentation sémantique de cette solution donnée par un expert est équivalente à la représentation sémantique correspondante à l'énoncé réellement prononcé* ». La représentation sémantique dont il est question ici se limite au sens utile de l'énoncé, tel que l'envisagent les approches sélectives de compréhension (cf. infra § 1.2) : Bousquet-Vernhettes C. (2002) Compréhension robuste de la parole spontanée dans le dialogue oral homme - machine : décodage conceptuel stochastique. Thèse de l'Université Paul Sabatier, Toulouse, France, 26 septembre 2002 (pages 72-73).

d'«*utile*»[6], qui se limite aux « *sujets directement liés à la tâche ou au canal de communication*[7] *»*. Ce que Wolfgang Minker résume par ces mots[8] :

> *«* [L'analyse] *doit se limiter aux éléments porteurs de sens de la requête* ***tout en ignorant les parties redondantes ou non-essentielles*** *pour l'application »* (souligné par nous)

La caractérisation de ce sens minimal est guidée par l'instanciation de schémas sémantiques prédéfinis qui regroupent un ensemble de segments conceptuels clés pour chaque type de requête. Considérons par exemple l'énoncé (4.1) :

(4.1) *Bonjour je voudrais les vols euh la liste des prochains vols directs pour Paris merci*.

L'analyse sélective ne retient que quelques bribes dans l'énoncé et leur associe le rôle qu'ils doivent y jouer (tableau 4.2, partie gauche). Ces segments conceptuels vont remplir le schéma pragmatico-sémantique représenté à droite du tableau.

Tableau 4.2 *— Compréhension sélective de la parole : exemple de schéma obtenu à partir d'une séquence de segments conceptuels*

request_category	*liste* (vols)	<type_requete = vols>
itinerary_stop	*direct*	<stop = NO>
itinerary_arrival	(pour) *Paris*	<from = LOCAL>
time_departure	*prochains*	<to = Paris>
		<date_departure = AUHOURD'HUI>
		<time_departure = MAINTENANT>

Ces principes étant posés, les méthodes sélectives ne se différencient plus à la marge que par l'algorithme d'analyse utilisé. Les systèmes reposent ainsi sur un décodage stochastique en segments conceptuels (systèmes CHRONUS ou CACAO[9]), l'utilisation d'automates ou règles probabilisés (système PHILIPS ou LIMSI ARISE-H[10]) ou encore des grammaires symboliques à base de cas sémantiques[11] (systèmes SUNDIAL, LATIS-R ou TRAINS[12]).

Dans tous les cas de figure, nous sommes en présence de traitements superficiels qui reposent au mieux sur des indices linguistiques très locaux pour remplir des schémas pragmatiques prédéfinis.

Le caractère local et partiel de ces traitements leur garantit une certaine robustesse en présence de parole spontanée. En dépit de leur pauvreté linguistique, ces approches ingénieriques ont ainsi démontré par le passé toute leur efficacité. Les systèmes sélectifs CHRONUS et PHOENIX sont par exemple arrivés aux deux premières places de la campagne d'évaluation ATIS (tableau 4.3).

---

6 Pérennou G. (1996) Compréhension du dialogue oral : le rôle du lexique dans l'approche par segments conceptuels. Actes de l'atelier *Lexique et Communication Parlée*, GDR-PRC CHM, Toulouse, France. 169-178.

7 Pierrel J-M., Romary L. (2000) Dialogue homme - machine. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 331-350 (citation page 332).

8 Minker W. (1999) Compréhension automatique de la parole. L'Harmattan, Paris, France (citation page 41).

9 Levin E., Pieraccini R. (1992) Chronus : the next generation. *Speech Communication*, 11, 283-288. ;Levin E., Pieraccini R. (1995) Concept-based spontaneous speech understanding. Actes *4th European Conference on Speech Communication and Technology*, *Eurospeech'95*, Madrid, Espagne. 555-558. ; Pieraccini R., Levin E. (1995) A spontaneous-speech understanding system for database query applications, Actes *ESCA Workshop on Spoken Dialogue Systems*. Vigso, Danemark. 85-88 ; Bousquet-Vernhettes C. (2002) *op. cit*.

10 Aust H., Oerder M., Seide F., Steinbiss V. (1995) The Phillips automatic train timetable information system. *Speech Communication*, 17. 249-26 ; Maynard H., Lefèvre F. (2002) Apprentissage d'un module stochastique de compréhension de la parole. Actes *XXIV° Journées d'Etudes sur la Parole, JEP'2002*, Nancy, France. 129-132.

11 C'est à Bruce qu'on doit la première mise en pratique effective de la théorie des cas sémantiques de Charles J. Fillmore : Bruce B. (1975) Case systems for natural language. *Artificial Intelligence*. 6. 327-360

12 Clementino D., Fissore L. (1993) A man-machine dialogue system for speech access to train timetable information. Actes *3rd European Conference on Speech Communication and Technology*. *Eurospeech'93*. Berlin, Allemagne. 1863-1866 ; Bennacef S. (1995) Modélisation du dialogue oral homme - machine : mise en œuvre dans une application de demande d'informations. Thèse de Doctorat, Université Paris XI, Orsay, France ; Allen J.F., Miller B.W., Ringger E.K., Sikorski T. (1996) Robust understanding in a dialogue system. Actes *34th Annual meeting of the Association for Computational Linguistics, ACL'96*. San Francisco, CA, 62-70).

Tableau 4.3 *— Taux de robustesse et type d'analyse mise en œuvre par les meilleurs systèmes de la campagne d'évaluation ARPA-ATIS de 1994[13]. Résultats obtenus sur des énoncés de type A (compréhension sans contexte)*

Laboratoire (Système)	AT&T (Chronus)	CMU (Phoenix)	MIT (Galaxy)	SRI (Gemini)	BBN (Hum)
% d'erreurs	3,8	3,8	4,5	7,0	9,4
Analyse	sélective	sélective	mixte	Profonde	sélective

Au vu de ces expérimentations, il est clair que ces méthodes sont parfaitement adaptées à des domaines très spécialisés (renseignement d'horaires de train, renseignement aérien, etc.). Il n'en reste pas moins que c'est le caractère finalisé de l'interaction qui leur permet de limiter la compréhension à la sélection superficielle de quelques concepts.

Or, la communication orale homme-machine oriente ses recherches vers des domaines d'application de plus en plus riches, quand elle ne s'intéresse pas à des applications multi-domaines[14]. Il me semble donc légitime de s'interroger sur les capacités d'extension des méthodes sélectives à des contextes moins restreints. En l'absence de campagnes d'évaluation sur ces domaines d'application plus complexes[15], on ne dispose que de peu d'éléments objectifs pour répondre à cette question. Je vais cependant tenter de cerner, sur quelques exemples, les problèmes que risquent de rencontrer à l'avenir les techniques sélectives de compréhension de la parole.

1.3. Les limites des méthodes sélectives : le problème de l'ambiguïté

Au cours des quinze dernières années, les recherches en dialogue homme-machine ont suivi une tendance continue à la complexification de la tâche considérée. Si l'on met de côté les recherches sur le dialogue multimodal[16], on est passé des systèmes de routage ou d'annuaire téléphonique[17] au domaine du renseignement aérien (programme DARPA-ATIS) ou ferroviaire (projet européen ARISE) et désormais au renseignement touristique (programme DARPA-Communicator, projet Verbmobil, campagne d'évaluation MEDIA).

Les analyses de corpus pilote que j'ai réalisées (cf. chapitre 2 § 3.1) ont confirmé qu'il n'existait pas de corrélation entre la richesse sémantique d'une tâche et la complexité structurelle des énoncés oraux rencontrés. Cette stabilité structurelle devrait militer en faveur de la généricité des méthodes sélectives vis-à-vis du domaine d'application. Ce serait oublier que la richesse sémantique d'une

---

13 Pallet D.S., Fiscus J.G. et al. (1995) 1994 benchmark tests for the ARPA spoken language program. *Actes 1995 ARPA workshop on spoken language technology*. Morgan Kaufman, Princeton, NJ. 5-36.

14 Chung G., Seneff S., Hetherington L. (1999) Towards Multi-Domain Speech Understanding Using a Two-Stage Recognizer, Actes *6th European Conference on Speech Communication and Technology*, *Eurospeech'1999*, Budapest, Hongrie. 2655-2658 ; Gufstafson J., Bell L. (2000) Speech technology on trial : experience from the August system. *Natural Language Engineering* , 6 (3-4). 273-286.
Cette problématique se retrouve au niveau de la reconnaissance de parole : Lefèvre F., Gauvain J.-L., Lamel L. (2002) Développement d'une technique générique pour la reconnaissance de la parole indépendante de la tâche. Actes *XXIV° Journées d'Etudes sur la Parole*, *JEP'2002*, Nancy, France, 221-224.

15 Voir le chapitre 3 consacré à l'évaluation des systèmes de dialogue.

16 Cette difficile problématique pose d'intéressantes questions en matière de fusions de modalités et, d'un point de vue plus linguistique, sur le calcul des références spatiales : Siroux J., Guyomard M., Jolly Y., Multon F., Remondeau C. (1995) Speech and tactile-based GEORAL system. Actes *3rd European Conference on Speech Communication and Technology* , *Eurospeech'95*, Madrid, Espagne, 1943-1946 ; Caelen J., Garcin P., Wretö J., Reynier E. (1991) Interaction multimodale autour de l'application ICP-Draw. Actes *IHM'91*. Dourdan, France. 1-12 ; Gaiffe B., Romary L., Pierrel J-M. (1991)) Reference in a multimodal dialogue: towards a unified processing. Actes *2nd European Conference on Speech Communication and Technology*, *Eurospeech'91*. Gènes, Italie.

17 Voir par exemple le système PJ (Pages Jaunes) développé conjointement par le CNET Lannion et l'IRISA : Guyomard M., Siroux J., Cozannet A. (1990) Le rôle du dialogue pour la reconnaissance de parole. Le cas du système Pages Jaunes. Actes *XVIII° Journées d'Etudes sur la Parole*, *JEP'1990*. Montréal, Canada. 322-326.

tâche se traduit également en terme d'ambiguïté sémantique. Or, c'est l'absence d'ambiguïté qui fonde la compréhension sélective[18].

Ce postulat semble devoir être remis en cause dans le cas d'applications complexes. Un seul exemple suffira à donner la mesure de ce problème. La tâche de renseignement touristique sur laquelle portent les travaux du groupe CORAIL comprend des termes qui peuvent recevoir jusqu'à 5 interprétations différentes[19]. A ma connaissance, une ambiguïté aussi élevée ne se rencontre pas sur des tâches très spécialisées comme le renseignement d'horaires de trains. Cela n'empêche pourtant pas les systèmes sélectifs actuels de rencontrer déjà de sérieuses difficultés en présence des rares cas d'ambiguïtés auxquels ils doivent faire face.

Dans une expérimentation portant sur l'interprétation mots inconnus ou mal reconnus, Caroline Bousquet montre par exemple qu'une « forte » ambiguïté sémantique — trois interprétations au maximum pour un même terme ! — peut avoir des conséquences critiques sur la compréhension sélective[20]. Elle observe en effet que la proportion d'énoncés mal compris par le système CACAO se situe entre 30 % et 85 % pour des mots ambigus, alors qu'il reste compris entre 18% et 32% en l'absence d'ambiguïté. Cette influence a également été constatée au cours de la campagne d'évaluation par défi que j'ai présentée au chapitre 3. On a observé à cette occasion que près de 14% des erreurs du système CACAO provenaient d'une mauvaise interprétation de la préposition ambiguë *à*. Avec pour conséquence une confusion entre les lieux de départ et d'arrivée[21].

Sans avoir force de démonstration, ces observations suggèrent que les approches sélectives ne sauraient être utilisées sur des domaines beaucoup plus riches que ceux actuellement envisagés. Certains chercheurs tels, Alex Waibel et Klaus Zechner, en arrivent à une conclusion identique[22]. Gerdjan van Noord conclut de même[23] :

> *« In [such extended applications], simple concept spotting may not be able to correctly process all constructions »*

D'autres éléments militent par ailleurs en faveur de solutions alternatives aux approches sélectives :

- L'effacement d'éléments jugés inutiles a des conséquences sur la compréhension en dehors du problème des éléments ambigus. Dans son mémoire de doctorat, Wolfgang Minker donne plusieurs exemples représentatifs d'erreurs dues à des effacements erronés[24]. Ces exemples sont révélateurs de la prise de risque sur laquelle se fonde la compréhension sélective.

---

18 Pierrel J-M., Romary L. (2000) *op. cit*.

19 Voir par exemple le cas du terme *horaire* dans le modèle de la tâche utilisé par le système ROMUS : Goulian J. (2002) Stratégie d'analyse détaillée pour la compréhension automatique robuste de la parole. Thèse Université de Bretagne Sud, Vannes, France. 13 Décembre 2002.

20 Bousquet-Vernhettes C. (2002) *op. cit.* (chapitre 4).

21 Cette proportion ne tient pas compte des erreurs imputables à un apprentissage insuffisant : Bousquet-Vernhettes C. (2002) *op. cit*. Données présentées en page 151 (figure 42).

22 Zechner K. (1998) Automatic construction of frame representations for spontaneous speech in unrestricted domains. Actes *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, *COLING-ACL'1998*. Montréal, Canada. 1448-1452.

23 van Noord G., Bouma G., Koeling R., Nederhof M.J. (1999) Robust grammatical analysis for spoken dialogue systems. *Natural Language Engineering*, 5(1).

24 Considérons l'énoncé attesté suivant donné par Minker (1999) *op. cit*. :

(4.2a) *après 19 heures j'ai une carte famille nombreuse*

Le système juge les séquences de mots *j'ai* et *carte* inutiles car elles ne relèvent pas du vocabulaire qui a été retenu pour décrire l'univers de la tâche. Ces mots « vides de sens » sont ignorés par la compréhension sélective. On se retrouve alors avec l'énoncé (4.2b) :

(4.2b) après **19 heures une** famille nombreuse

La compréhension sélective ne procédant qu'à une analyse lexicale locale, le déterminant *une* est mal interprété. La séquence de mots mise en relief est alors considérée à tort comme un horaire : 19h01 !

- En se limitant à une analyse locale et superficielle des énoncés (*concept spotting*), la compréhension sélective rencontre des difficultés à traiter des structures complexes telles que les négations[25], les requêtes multiples[26] ou encore les relations prédicat / arguments récursives.
- Les approches sélectives tendent à simplifier l'identification des informations nécessaires à la gestion du dialogue. Certes, les schémas sémantiques qu'elles utilisent comportent une mention de l'acte de dialogue (confirmation, requête, information, etc.) porté par l'énoncé. La caractérisation de l'acte repose toutefois sur l'identification grossière de quelques mots clés. En particulier, les marqueurs linguistiques qu'utilisent Nathalie Colineau et Jean Caelen[27] pour la détermination fine des actes de dialogue sont généralement éliminés par une compréhension sélective. Leur désambiguïsation[28] nécessite par ailleurs une analyse contextuelle qui ne saurait être réalisée par des traitements purement locaux.

Ces observations me conduisent à penser qu'il y a une place en compréhension de le parole pour des traitements linguistique plus profonds. Bien entendu, renoncer aux méthodes sélectives revient à s'exposer de front aux problèmes de robustesse.

En 1994, Renato De Mori refusait toute alternative aux approches sélectives en ces termes[29] :

> *« L'approche purement syntaxique tend à se décomposer lors de la présence de mots inconnus, de nouvelles constructions linguistiques, d'erreurs dans la reconnaissance, et de phénomènes spontanés dans la parole comme les répétitions »*

Dix ans plus tard, une analyse linguistique fine de la parole spontanée est-elle toujours impossible ? Van Noord et ses collègues sont d'un avis opposé lorsqu'ils affirment qu'une analyse syntaxique détaillée est la seule alternative envisageable sur des domaines d'application plus complexes[30] :

> *« The grammatical approach may become essential as soon as the application is extended »*

C'est ce que je vais montrer à la lumière des travaux de notre équipe. Avant de présenter en détail les systèmes de compréhension de parole réalisés au VALORIA, je vais cependant revenir sur des recherches présentant certaines affinités avec notre approche.

1.4. Traitements linguistiques fins pour une compréhension de parole complexe

Plusieurs laboratoires ont déjà recherché à fonder la compréhension de la parole sur une analyse linguistique profonde. Le premier système qu'il convient d'étudier est TINA qui a été développé au MIT par Stéphanie Seneff et ses collègues[31]. TINA est en réalité un système hybride :

- d'une part, un module linguistique procède à une analyse hors contexte de la structure profonde de l'énoncé. La grammaire utilisée est transformée par apprentissage en un automate

---

[25] Considérons l'exemple (4 ?3a) que l'on doit également à Minker (1999) *op. cit* :

(4.3a) *je ne pars pas de Montpellier le 17 décembre je voudrais partir le 11 décembre de Montpellier*

Suite à l'élimination des mots jugés inutiles, l'énoncé devient :

(4.3a) *pars pas de Montpellier 17 décembre partir 11 décembre de Montpellier*

Quoique lapidaire, cet énoncé reste compréhensible. L'absence d'analyse profonde de l'énoncé va pourtant conduire à une confusion sur la portée de la négation. Celle-ci ne concernera pas la date de départ, mais la ville. Pour le système, l'utilisateur demande effectivement à partir le 17 décembre, mais d'une autre localité !

[26] Minker W., Bennacef S. (1996) Compréhension et évaluation dans le domaine ATIS. Actes *XXI° Journées d'Etudes sur la Parole*, *JEP'1996*, Avignon, France. 415-419.

[27] Colineau N., Caelen J. (1997) Analyses de dialogues oraux et modélisation des actions de communication. Actes des *1ères Journées Scientifiques et Techniques FRANCIL*, *JST'1997*, Avignon, France, 447-454.

[28] Les auteurs donnent par exemple le cas du marqueur *voilà*, qui peut caractériser une confirmation, un acquiescement, ou une clôture de tâche mais peut être également utilisé comme phatique ou présentatif : Colineau N., Caelen J. (1997) *op. cit*. (p. 448).

[29] De Mori R. (1994) Apprentissage automatique pour l'interprétation sémantique. Actes des *XX° Journées d'Etudes sur la Parole*, *JEP'1994*. Trégastel, France. 11-19 (citation p. 13).

[30] van Noord G., Bouma G., Koeling R., Nederhof M.J. (1999) *op. cit*.

[31] Seneff S. (1992) TINA: a natural language system for spoken language applications. *Computational Linguistics*, 18(1). 61-86

probabiliste. Les nœuds de l'automate peuvent correspondre aussi bien à des catégories syntaxiques que sémantiques (figure 4.1). La grammaire est donc spécifique de l'application étudiée.

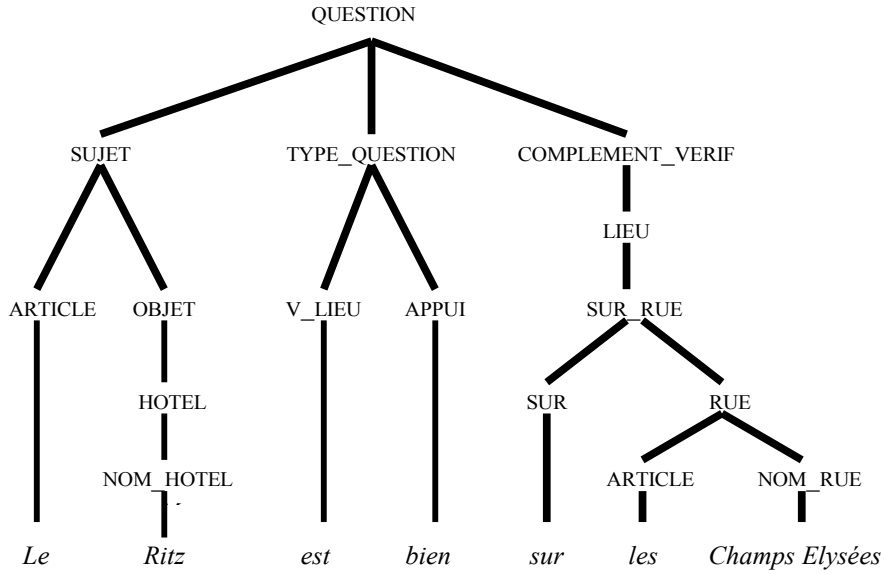

Figure 4.1 *— Exemple de représentation sémantique détaillée fournie par le système TINA*

- d'autre part, un module de compréhension sélective (effacement de certains mots inutiles, caractérisation de concepts, etc.) prend le relais lorsque l'analyse fine a échoué[32].

La campagne d'évaluation ATIS a montré la pertinence de cette architecture. Le système de dialogue GALAXY, qui utilise TINA, a présenté un taux d'erreur de phrases proche de celui des meilleurs systèmes sélectifs- (tableau 4.3). Pour pouvoir apprécier ce résultat à sa juste valeur, il conviendrait néanmoins de savoir dans quelles proportions le module sélectif a pris le relais de l'analyse linguistique détaillée. De l'aveu même des concepteurs du système, il semble ainsi que le module linguistique de TINA ne soit pas nécessaire pour des tâches simples telles que l'information météorologique (application JUPITER[33]).

Le système JANUS, développé par Alex Waibel conjointement à CMU et à l'Université de Karlsruhe, repose également sur une stratégie hybride combinant compréhension sélective (module PHOENIX) et détaillée[34]. Ici, l'analyseur syntaxique utilise une grammaire lexicale fonctionnelle (LFG).

Du côté francophone, Patrice Lopez (LORIA) propose d'analyser les énoncés oraux à l'aide d'une grammaire d'arbres adjoints lexicalisés[35]. Les LTAG ont déjà donné lieu à des tentatives d'analyse syntaxique de surface (*shallow parsing*) robuste[36]. Ici, elles sont utilisées de manière profonde suivant une stratégie par expansion d'îlots qui autorise une recherche partielle en cas de difficultés. Le comportement du système semble relativement robuste. Ce système n'a malheureusement pas participé à la campagne d'évaluation par défi, ce qui aurait permis de se faire une idée plus objective sur la pertinence de cette proposition.

---

32 Seneff S. (1992) Robust parsing for spoken language systems. Actes *International Conference on Acoustics, Speech and Signal*, *ICASSP'1992*. San Francisco, Etats-Unis. 189-192.

33 Glass J. (1999) Challenges for spoken dialogue systems. Actes *IEEE ASRU Workshop*. Keystone, Colorado, Etats-Unis.

34 JANUS est en fait le nom du module de reconnaissance de la parole sur lequel est basé le système : Waibel A. (1996) Interactive translation of conversational speech. *Computer*, 27(7). 41-48.

35 Lopez P. (1999) Représenter et utiliser les contraintes de la langue oral à l'aide d'une grammaire lexicalisée d'arbres adjoints. Actes *TALN'99*, Cargèse, France. 445-450.

36 Bangalore S. (1999) Suppertagging : an approach to almost parsing. *Computational Linguistics*, 25(2). 237-265.

Cet état de l'art montre la viabilité d'une compréhension reposant sur des traitements linguistiques fins. Les systèmes que je viens de présenter répondent à une même approche[37]. Ils mettent tous en jeu une analyse syntaxique profonde (CFG, LFG, TAG), les problèmes de robustesse sur l'oral spontané étant résolus par une stratégie d'analyse partielle ou le recours à une architecture hybride.

A l'opposé, j'ai orienté les travaux de notre équipe sur une approche alternative qui s'inspire d'un ensemble de recherches que l'on regroupe généralement sous le terme de TAL robuste (*shallow parsing*). Après avoir justifié le choix de cette approche originale, je présenterai en détail son adaptation à la problématique de la compréhension de la parole.

1.5. Les réalisations du VALORIA

Avant de présenter les systèmes de compréhension réalisés au VALORIA, il me semble utile de revenir sur mes travaux de doctorat. Ce retour sur des recherches déjà anciennes éclairera en effet les choix effectués depuis.

1.5.1. Une première esquisse bien peu ingénierique : le système ALPES

Le système ALPES[38], développé entre 1991 et 1994 à l'Institut de la Communication Parlée puis au CLIPS-IMAG, visait déjà une compréhension fine de la parole spontanée.

Le système ALPES reposait sur trois principes fondateurs :

- La représentation sémantique des énoncés ne se limitait pas à un simple schéma attributs / valeurs comme dans les systèmes sélectifs, mais à une structure de traits pouvant rendre compte de dépendances profondes.
- Les rôles sémantiques considérés ne correspondaient pas à des cas pragmatiques orientés vers la tâche[39] (Tarif, Départ, etc.), mais à des cas sémantiques génériques (Agent, Objet, etc…) semblables à ceux définis par Charles Fillmore[40].
- L'analyse était fondée sur des considérations purement sémantiques dans l'espoir de contourner les problèmes de robustesse rencontrés par les parseurs syntaxiques. L'idée était de caractériser des dépendances entre les mots de l'énoncé par la recherche de compatibilités entre leurs traits sémantiques (sèmes). Cette analyse se basait sur le substrat théorique de la sémantique interprétative de François Rastier[41].

Le lexique sémantique décrivant chaque lexème et les relations que peuvent partager ses sèmes était compilé sous la forme d'un réseau d'amorçage sémantique. L'énoncé est traité suivant une analyse gauche droite. Celle-ci consistait à relier les mots suivant les dépendances lexicales caractérisées par le processus d'amorçage. Les expérimentations réalisées sur le corpus ICP-Draw[42] ont montré la robustesse de cet analyseur détaillé sur le domaine de la conception de dessin.

---

37 Le cas du système de compréhension du LIMSI, développé par Sophie Rosset, est particulier. La majeure partie de l'interprétation est en effet déléguée à un contrôleur de dialogue très fin, le module de compréhension se limitant à une pré-analyse limitée de type sélectif. Cette mise en avant de l'interprétation contextuelle est intéressante. On peut par contre s'interroger sur les limites de l'analyse littérale initiale : Rosset S. (2000) Stratégies et gestionnaire de dialogue pour des systèmes d'interrogation de bases de données à reconnaissance vocale. Doctorat Université Paris XI, Orsay, France. Publié comme rapport de recherche 2001-18 du LIMSI-CNRS, Orsay, France. septembre 2001

Je n'ai pas retenu non plus le système OASIS développé par Mohamed Zakaria Kurdi au CLIPS-IMAG. En dépit des affirmations de son concepteur, le formalisme STAG utilisé dans OASIS relève d'une approche sélective : Kurdi M.Z. (2003) Analyse linguistique robuste et profonde du langage oral spontané. Thèse Université Joseph Fourier, Grenoble, France

38 ALPES : Analyse Linguistique de la Parole en Elocution Spontanée : Antoine J.Y. (1994) Coopération syntaxe -sémantique pour la compréhension de la parole spontanée. Thèse de Doctorat. INP Grenoble, Grenoble, France.

39 Bruce B. (1975) *op. cit*.

40 Fillmore C. J. (1968) The case for case. *In* Bach E., Harms R. (Eds) *Universals in Linguistic Theory*. Holt et Rinehart and Winston, New-York, Etats-Unis. 1-90.

41 Rastier F. (1987) Sémantique interprétative. PUF, Paris, France.

42 Caelen J., Garcin P., Wretö J., Reynier E. (1991) Interaction multimodale autour de l'application ICP-Draw. Actes *IHM'91*. Dourdan, France. 1-12.

Les motivations qui ont guidé ce travail de doctorat me semblent toujours légitimes. Cependant, le système ALPES présentait des insuffisances qui obéraient sa généralisation à d'autres domaines d'applicatifs :

- D'un point de vue ingénierique, l'effort de constitution d'un lexique reposant sur une taxonomie de sèmes — 3 mois d'analyse sémantique pour un lexique de 200 mots — s'est avérée être un frein considérable à la portabilité du système.
- D'un point de vue plus théorique, le recours à une analyse exclusivement sémantique induit des limitations que j'avais relevées sur les approches sélectives. Dès que l'on quitte des applications très restreintes, le problème de l'ambiguïté ne peut en effet plus être résolu par une analyse purement sémantique (ou pragmatique). Cette observation fut manifeste avec ALPES, toute augmentation de la taille du lexique se traduisant par une explosion de la combinatoire d'analyse.

Sans renier certains principes tels que la recherche des dépendances sémantiques profondes de l'énoncé, cette expérience m'a conduit à revaloriser le rôle structurant de la syntaxe. D'où mes recherches actuelles, qui cherchent à utiliser aussi loin que possible des traitements syntaxiques robustes avant de faire appel à une connaissance sémantico-pragmatique.

1.5.2. TAL robuste et ingénierie des langues

Cette stratégie d'analyse incrémentale s'inspire d'un courant en pleine émergence de l'ingénierie des langues qui est qualifié généralement de « TAL robuste » (*robust parsing*). Historiquement, le TAL robuste a constitué une des premières tentatives concluantes de concilier robustesse d'analyse et TALN. Il a conduit dès le milieu des années 1990 à la mise en œuvre d'analyseurs syntaxiques robustes du langage écrit[43].

Ces analyseurs robustes reposaient sur trois postulats principaux :

- Ils mettent en œuvre une analyse complète mais superficielle (*shallow parsing*) des énoncés. En particulier, l'analyse peut se limiter à une segmentation en constituants minimaux non récursifs, appelés segments noyaux ou encore *chunks*[44].
- L'analyse est sous-spécifiée. En particulier, la robustesse est privilégiée à une désambiguïsation complète — mais risquée — à un niveau d'analyse donné. Si l'analyseur ne dispose pas des connaissances nécessaires pour trancher à coup sûr entre deux alternatives, il conserve ces deux solutions en réponse.
- Conséquence du point précédent, l'analyse est non destructrice c'est-à-dire qu'elle conserve toute l'information présente dans l'énoncé pour d'éventuels traitements ultérieurs. On évite ainsi une transmission en chaîne des erreurs issues d'un module d'analyse particulier.

Au cours des dernières années, cette démarche a été généralisée à plusieurs niveaux de traitements. On construit ainsi des analyseurs syntaxiques profonds et robustes qui reposent sur une analyse incrémentale où chacune des étapes successives de traitement répond aux mêmes exigences[45] :

- A chaque niveau, l'analyse est sous-spécifiée, non destructrice et superficielle, au sens où la profondeur des représentations élaborées ne s'accroît que légèrement d'un niveau à l'autre,
- Chaque étape est conceptuellement indépendante de l'analyse globale. C'est-à-dire que les connaissances manipulées à un niveau donné font sens par elles-mêmes. Cette cohérence interne facilite la conception et la maintenance des systèmes, ainsi que la correction des erreurs.
- La décomposition modulaire de la connaissance permet l'utilisation de techniques d'analyses

---

43 Ejerhed E. (1993) Nouveaux courants en analyse syntaxique, *TAL*, 34(1), 61-82.

44 Abney S. (1991) Parsing by chunks. *In* Berwick R., Abney S. and Tenny C. (Eds.) *Principle-based parsing*. Kluwer Academic Publ., Dordrecht, Pays-Bas. Disponible sur la Toile : www.sfs.nphil.uni-tuebingen.de/~abney.

45 Ait-Mokhtar S., Chanod J.-P., Roux C. (2002) Robustness beyond shallowness : incremental deep parsing. *Natural Language Engineering*, 8(2-3). 121-144 ; Basili R., Zanzotto F.M (2003) Parsing engineering and empirical robustness. *Natural Language Engineering*, 8 (2-3) 147-169.

efficaces. Le TAL robuste se caractérise en particulier par l'utilisation fréquente d'automates à états finis, le pouvoir génératif des grammaires hors contexte n'étant plus nécessairement requis pour un niveau de traitement donné.

La notion d'analyse incrémentale peut également s'appliquer à la structure interne des modules de traitement. A chaque niveau, l'analyse est alors conduite en plusieurs passes successives grâce à la définition d'une hiérarchie de règles (ou d'automates) qui s'appliquent les unes après les autres. On en arrive à une décomposition descriptive de la connaissance manipulée. Par exemple :

- on définit tout d'abord un ensemble de règles générales qui ne décrivent qu'une partie de la réalité et ne s'appliquent que quand elles le peuvent,
- d'autres règles sont ensuite appliquées, qui rendent compte d'un ensemble de cas particuliers,
- Enfin, un dernier groupe de règles de moins contraintes est activé lorsque les passes précédentes n'ont pas conduit à une analyse concluante (énoncés fortement perturbés).

Jusqu'à présent, ces approches ont surtout été utilisées pour le traitement de textes libres (langage écrit) où elles ont largement fait la preuve de leur efficacité et de leur robustesse. L'originalité des travaux que je vais présenter réside précisément dans leur adaptation au langage parlé et plus précisément à la compréhension de parole spontanée[46].

1.5.3. Les systèmes de compréhension de la parole du laboratoire VALORIA

Deux systèmes de compréhension ont été développés au VALORIA dans le cadre de doctorats effectués sous ma direction scientifique[47]. Il s'agit respectivement du système ROMUS de Jérôme Goulian et du système LOGUS de Jeanne Villaneau. A l'heure actuelle, ROMUS ne réalise qu'une compréhension littérale des énoncés, par opposition au système LOGUS qui intègre une résolution des co-références anaphoriques. L'objectif de ces deux systèmes est de réaliser une compréhension fine et profonde des énoncés oraux. Ils fournissent ainsi en sortie une représentation sémantique qui correspond à un arbre (ou un graphe) de dépendances entre les mots de l'énoncé. Suivant le système, il s'agit d'une structure de traits récursive (ROMUS) ou d'une formule logique à la Montague (LOGUS). Le domaine d'application retenu est le renseignement touristique.

Ces deux systèmes reposent sur une stratégie d'analyse incrémentale qui comporte quatre niveaux de traitements principaux (figure 4.2). A l'exception de la reconnaissance de la parole, chaque étape met en œuvre une analyse superficielle qui répond aux principes du TAL robuste.

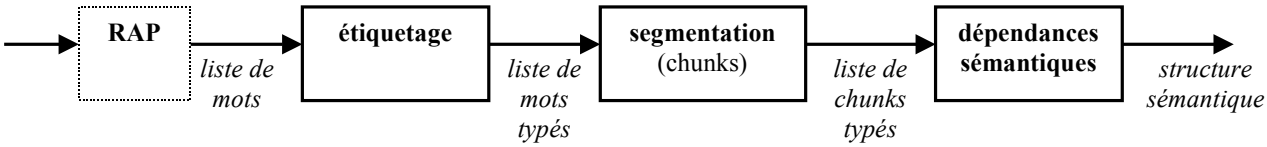

Figure 4.2 *— Architecture générique des systèmes de compréhension LOGUS et ROMUS*

Reconnaissance automatique de la parole (RAP) — Le groupe CORAIL ne mène pas de recherches spécifiques dans le domaine de la reconnaissance de la parole. Nous utilisons donc des systèmes de dictée vocale grand public en entrée de la chaîne de traitement décrite ci-dessus. Cette stratégie a ses limites. En particulier, les systèmes de dictée vocale ne sont pas adaptés au dialogue interactif. C'est pourquoi nous avons prévu de procéder prochainement à des expérimentations utilisant les sorties d'un système de reconnaissance issu de la recherche scientifique. Cette expérience sera réalisée en collaboration avec le laboratoire IRISA/Cordial et concernera le système

[46] Les travaux de Klaus Zechner (CMU et U. Karlsruhe), suivent une approche similaire à la notre : Zechner K. (1998) Automatic construction of frame representations for spontaneous speech in unrestricted domains. Actes *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, *COLING-ACL'1998*. Montréal, Canada. 1448-1452.

[47] Le doctorat de Jeanne Villaneau fait l'objet d'un co-encadrement avec Olivier Ridoux (IRISA, Rennes). Franck Poirier (VALORIA, UBS) était le directeur de thèse officiel de Jérôme Goulian.

LOGUS. Par ailleurs, nous espérons également pouvoir utiliser des sorties réelles de système de reconnaissance dans le cadre de la campagne d'évaluation MEDIA.

Etiquetage — Cette étape consiste à associer une étiquette morphosyntaxique à chaque mot de l'énoncé. Plus précisément, on utilise un jeu de catégories grammaticales (parties du discours), auquel s'ajoute un nombre limité d'informations morphologiques. A la différence des étiqueteurs morphosyntaxiques utilisés à l'écrit, cette étape ne vise qu'une désambiguïsation limitée. Cette sous-spécification nous permet d'atteindre une précision d'étiquetage élevée, cruciale sur l'oral spontané[48], au détriment d'une faible décision. Il apparaît que dans le cas du renseignement touristique, le degré d'ambiguïté morphosyntaxique reste suffisamment limité pour autoriser une analyse parallèle des différentes hypothèses d'étiquetage. Au final, plusieurs hypothèses d'étiquetage peuvent donc être proposées au processus de segmentation.

Segmentation — Cette étape consiste à segmenter l'énoncé en une liste de constituants (groupes nominaux, verbaux, adjectivaux, etc.) minimaux non récursifs. Ces segments noyaux recoupent la notion de *chunks* définie par Steven Abney[49]. La segmentation est fondée sur une connaissance syntaxique qui porte sur les parties du discours caractérisées par l'étape précédente. Cette analyse de surface présente plusieurs intérêts dans la perspective d'une compréhension fine et robuste :

- les chunks correspondent généralement à des unités de sens représentant les objets de la tâche. La segmentation facilite donc la transition vers la représentation sémantique de l'énoncé,
- le caractère superficiel de la segmentation en chunks permet d'atteindre une certaine robustesse[50] tout en conduisant une analyse plus détaillée que les méthodes sélectives. Aucun élément de l'énoncé n'est en effet ignoré à ce stade,
- le chunk est une unité de segmentation adaptée au langage parlé spontané. Il a en effet été démontré que ces constituants minimaux sont le lieu de réalisation privilégié des réparations à l'oral. Comme le rappelle Claire Blanche-Benveniste [51] :

  « pour corriger un élément, on reprend généralement depuis le début du syntagme »

  Cette observation permet d'envisager un traitement intégré des réparations en analysant le *reparandum* et l'altération comme la réalisation de deux *chunks* distincts. Afin d'aligner la segmentation avec ces éléments oraux, nos *chunks* ont une taille minimale comme le montre la figure 4.3 qui correspond à la segmentation de l'énoncé suivant :

  (4.3) *je veux les tarifs pour une chambre enfin une chambre double à l'hôtel Caumartin et au Crillon*.

---

48 André Valli et Jean Véronis ont montré que l'étiquetage morphosyntaxique de la parole spontanée pouvait être relativement robuste. Les taux d'erreurs rapportés restent toutefois trop importants pour faire reposer l'analyse sur une séquence d'étiquettes désambiguïsées : Valli A. et Véronis J. (1999) Etiquetage grammatical de corpus de parole : problèmes et perspectives. *Revue Française de Linguistique Appliquée*, 4(2), 113-133.

49 Abney S. (1991) Parsing by chunks. *In* Berwick R., Abney S., Tenny S. (Eds.) *Principle based parsing*. Kluwer Academic Publ., Dordrecht, Pays-Bas. Disponible sur la Toile : www.sfs.nphil.uni-tuebingen.de/~abney.

50 Cette robustesse a en tous cas été observée pour de nombreux segmenteurs travaillant sur du texte écrit : Church K. (1988). A stochastic parts program and noun phrase parser for unrestricted text, actes *Conference on Applied Natural Language Processing*, *ACL'1988*, Austin, TX, 136-143 ; Koskiennnemi K. (1990) Finite state parsing and disambiguation. Actes *13th Conference on Computational Linguistics*, *COLING'90*. Helsinki, Finlande. 229-232 ; Chanod J.P. (1994) Developpements en analyse syntaxique automatique. Actes TALN'1994, Marseille. France. 87-91;
L'analyseur du GREYC, qui a fait montre de sa robustesse comme étiqueteur au cours de la campagne d'évaluation GRACE, repose également sur une segmentation en chunks : Vergnes J., Giguet E. (1998) Regards théoriques sur le tagging. Actes *TALN'1998*, Paris, France. 22-31.
La prochaine campagne EASyY d'évaluation des analyseurs syntaxiques du français (action TECHNOLANGUE) portera entre autre sur la segmentation en chunks des énoncés. Le corpus de test comprendra de la parole spontanée transcrite, ce qui permettra ainsi d'avoir une idée plus précise de la robustesse de ces méthodes sur le langage parlé.

51 Blanche-Benveniste C. (1997) Approches de la langue parlée en français, Coll. *L'essentiel Français*, Ophrys, Paris, France (p. 47 – 49) ; Voir aussi : Martinie B. (2001) Remarques sur la syntaxe des énoncés réparés en français parlé. *Recherches sur le Français Parlé*, 16 (2001), 189-206.

[*je*] [*veux*] [*les tarifs*] [*pour une chambre*] [*enfin*] [*une chambre double*] [*à l'hôtel*] [*Caumartin*] [*et*] [*au Crillon*]

Figure 4.3 *— Exemple de segmentation réalisée par les systèmes LOGUS et ROMUS.*

- la segmentation n'est pas destructrice en présence de réparations. Contrairement aux techniques de pré-traitements par normalisation, le reparandum n'est pas effacé par l'altération[52]. On conserve ainsi l'ensemble de l'information véhiculée par le message oral. L'intérêt de cette préservation est évident dans les cas d'enrichissement lexical. Il peut être également utile lorsque l'altération présente des erreurs de reconnaissance qui empêchent son interprétation. Les réparations seront analysées lors de l'étape de caractérisation de dépendances entre *chunks*.
- la segmentation repose comme l'étiquetage sur une connaissance syntaxique indépendante de la tâche. Ces deux premières étapes de traitement présentent donc d'indéniables atouts en matière de généricité. Ce ne sera pas le cas de l'étape de caractérisation des dépendances entre chunks.

Caractérisation des dépendances — La segmentation permet une caractérisation des dépendances internes à chaque chunk. La dernière étape est au contraire en charge de la détection des dépendances entre ces segments noyaux, afin d'élaborer la structure complète de l'énoncé.

En règle générale, les relations de dépendances associent les têtes lexicales des chunks. Dans un premier temps, chaque constituant syntaxique est traduit en un segment conceptuel par consultation d'un lexique spécifique à l'application. Ce lexique pragmatique décrit les attentes des concepts sous la forme de relations prédicat / argument entre les objets de la tâche (par exemple, un tarif peut être associé à différentes propriétés).

L'analyse revient à construire un graphe de dépendances par recherche d'associations. Elle suit une heuristique privilégiant la solution la plus couvrante qui minimise le nombre des concepts finaux ou, suivant le système, les dépendances les plus courtes. On autorise également une analyse partielle (relâchement de contraintes) en cas de difficultés. De même, les réparations sont identifiées à ce niveau. Les chunks qui correspondent au *reparandum* et à l'*altération* remplissent en effet le même rôle pragmatique. Une fois détectés, ils sont associés dans un même concept.

Cette étape cherche également à déterminer l'acte de dialogue véhiculé par l'énoncé. Suivant le principe de sous-spécification évoqué plus haut, cette caractérisation reste volontairement grossière (*assertion* / *demande* / *confirmation*, etc.). Le contrôleur de dialogue l'affinera ultérieurement.

En conclusion, ce niveau de traitement met en œuvre une analyse guidée par la tâche que l'on retrouve en première approximation dans les systèmes sélectifs. Notre démarche reste cependant originale à deux points de vue :

- l'analyse des dépendances concerne l'ensemble des chunks, et non pas certains segments clés.
- elle intervient uniquement en fin d'analyse incrémentale. Les niveaux de traitement inférieurs — plus originaux dans le cadre de la compréhension de la parole — confèrent au contraire aux systèmes ROMUS et LOGUS des propriétés intéressantes en terme de généricité, de robustesse et de finesse d'analyse.

Je vais précisément m'attarder maintenant sur les caractéristiques de ces systèmes, avant de revenir sur l'analyse de leur comportement au vu de plusieurs expérimentations.

[52] cf. chapitre 2, § 3.3.

1.5.4. Système ROMUS

ROMUS[53] est certainement celui des deux systèmes qui répond le plus fidèlement à l'architecture décrite ci-dessus. Afin de mieux comprendre le fonctionnement du système, je donnerai toutefois quelques informations supplémentaires sur son implémentation.

Etiquetage — Le lexique utilisé pour l'étiquetage en parties du discours (45 000 formes fléchies, 34 étiquettes grammaticales dont 6 portant une distinction singulier/pluriel) est représenté sous la forme d'un automate à états finis déterministe. La désambiguïsation s'effectue à l'aide d'un nombre limité de règles locales compilées en transducteurs déterministes utilisés en cascade. Plus précisément, cinq transducteurs sont appliqués successivement, chacun d'entre eux encodant plusieurs règles dont les contextes d'application ne se recouvrent pas (analyse incrémentale). La désambiguïsation reste limitée afin d'éviter des erreurs pénalisantes pour la suite de l'analyse. L'étiquetage présente ainsi un taux de décision de 80,4 % (corpus de test de 1200 énoncés) pour une précision de 97,5 %.

Le tableau 4.4 (page suivante) donne la liste des étiquettes grammaticales retenues par le système ROMUS. On relèvera leur caractère purement syntaxique qui nous assure la généricité de l'analyse. Cette propriété se retrouve dans l'étape suivante de segmentation.

Segmentation — La segmentation repose sur une modélisation symbolique. Chaque *chunk* est décrit par un ensemble d'expressions régulières portant sur les parties du discours (étiquettes grammaticales) des mots de l'énoncé. Ces expressions sont compilées en transducteurs que l'on rend déterministes[54]. Chaque transducteur est utilisé en cascade pour introduire dans l'énoncé des marqueurs de délimitation autour des segments caractérisés. L'ambiguïté de la segmentation est gérée par une heuristique de maximisation des segments détectés. Au final, trois types de chunks peuvent être caractérisés :

- constituants minimaux non récursifs généraux (*chunks* nominaux, verbaux, adjectivaux etc.),
- segments correspondant à des expressions langagières particulières (date, heure, prix),
- marqueurs de l'oral spontané (interjections, appuis du discours, etc.).

***Tableau 4.4** — Jeu d'étiquettes grammaticales retenu par le système ROMUS*

Adj	adjectif	**Vpp**	verbe au participe passé
Adjint	adjectif interrogatif	**Auxpp**	auxiliaire au participe passé
Num	numéral	**Modpp**	modal au participe passé
Adv	adverbe	**Vpr**	verbe au participe présent
Advq	adverbe de quantité	**Auxpr**	auxiliaire au participe présent
Advqs	adv. marqueur de superlatif	**Modpr**	modal au participe présent
Neg	discordantiel	**Nomc**	nom commun
Conjcoo	conjonction de coordination	**Nomp**	nom propre
Artind[sg/pl]	article indéfini	**Ppers**[sg/pl]	pronom personnel
Artdef[sg/pl]	article défini	**Pdem**[sg/pl]	pronom démonstratif
Adjpos[sg/pl]	adjectif possessif	**Pref**	pronom réfléchi
Adjdem[sg/pl]	adjectif démonstratif	**Prel**	pronom relatif
Prep	préposition	**Pint**	pronom interrogatif

[53] Pour une présentation récente du système ROMUS (ROBust Message Understanding System), on consultera la thèse de Jérôme Goulian : Goulian J. (2002) *op. cit.* ; Goulian J., Antoine J-Y., Poirier F. (2002) Compréhension automatique de la parole et TAL : une approche syntaxico-sémantique pour le traitement des inattendus structuraux du français. Actes *TALN'2002*, Nancy, France. vol. 1, 389-394.

[54] Roche E., Schabes Y (1997) Finite state language processing, MIT Press, Cambridge, MA (chapitre Deterministic Part of Speech Tagging with Finite State Transducers. 205-239)

Vinf	verbe à l'infinitif	**Conjsub**	conjonction subordination
V	verbe conjugué	**Conjsub**	conjonction subordination
Aux	auxiliaire conjugué	**X**	phatique
Mod	verbe modal conjugué		

Les règles de segmentation permettent un typage syntaxique des constituants ainsi qu'une caractérisation de leur structure (tête lexicale, dépendances locales).

Les mots qui ne sont pas intégrés à un segment à l'issue de la segmentation sont éliminés. Contrairement aux approches sélectives, ces rejets sont très limités. En règle générale, il s'agit d'amorces de réparations non porteuses de sens. A titre d'illustration, la figure 4.4 donne la segmentation effectuée par ROMUS de l'énoncé (4.5), qui comporte une reprise :

(4.5) *je cherche un un restaurant euh un restaurant chinois*.

$[je*]_{PR}$ $[cherche*]_{GV}$ ***un*** $[un\ restaurant*]_{GN}$ $[euh]_{HES}$ $[un\ restaurant*]_{GN}$ $[chinois*]_{GADJ}$

Figure 4.4 *— Segmentation de l'énoncé « je cherche un un restaurant euh un restaurant chinois » par le système ROMUS. La tête lexicale des segments est marquée par un astérisque.*

On observe que le premier déterminant *un* qui est répété en amorce du reparandum, sera éliminé. A l'opposé, le reparandum lui-même (*un restaurant*) est bien conservé, de même que le terme d'édition *euh*. Comme je l'ai évoqué plus haut (cf. 1.5.5), l'analyse considère des chunks réellement minimaux afin de coller aux unités de réalisation des réparations orales. Cela explique que, sur notre exemple, l'adjectif postposé *chinois* n'est pas intégré au groupe verbal qu'il qualifie[55].

En fin de segmentation, chaque *chunk* est représenté par un arbre dont la racine est un triplet <*S*, *T*, *M*>, où *S* est la catégorie syntaxique du segment, *T* sa tête lexicale et *M* un ensemble de marques morphologiques (pluriel / singulier, défini / indéfini).

Dépendances — La caractérisation des dépendances entre *chunks* repose sur le formalisme des grammaires de liens défini par Sleator et Temperley[56]. Comme nous l'avons vu au chapitre 2 (cf. § 3.2), les grammaires de liens ne peuvent modéliser les structures non projectives mais nos études de corpus pilotes ont montré que cette limitation est sans conséquence en dialogue finalisé.

Nous avons adapté le formalisme à la problématique du dialogue finalisé afin de permettre la manipulation d'une connaissance sémantico-pragmatique. A chaque entrée de la grammaire de liens (lexique) correspond un ensemble d'attentes sémantiques spécifiques à la tâche. Elles s'expriment au moyen de connecteurs étiquetés et orientés qui doivent s'associer deux à deux pour former une relation valide. Ces relations concernent des triplets *<S,T,M>*[57] et non pas des mots comme dans le formalisme originel. Deux types de relations ont été définis :

- relations spécifiques aux items lexicaux, comme par exemple la relation *Catégorie* qui peut relier deux segments <*GN, « hôtel », indéfini*> et <*GN, « étoile », défini*>,
- relations exprimant des constructions syntaxiques génériques comme les coordinations marquées par une conjonction (<*Coo, «et», ∅*> par exemple). Les réparations marquées par un terme d'édition explicite sont également modélisées ainsi. Si l'on reprend l'exemple de l'énoncé (4.3), la réparation sera identifiée par la présence du terme d'édition *enfin* qui est défini, dans une de ses acceptations, comme une conjonction : <*Cor, « enfin », ∅*>.

[55] Remarquons qu'une convention analogue a été adoptée dans le cadre de la campagne EASy d'évaluation des analyseurs et segmenteurs syntaxiques.

[56] Sleator D. D. K., Temperley D. (1991) Parsing English with a link grammar , *rapport de recherche CMU-CS-91-196*, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA.

[57] En réalité, l'élément S du triplet est transformé en un rôle sémantique RS vis-à-vis de l'application (objet, localisation, etc…).

Le dictionnaire comporte environ 1000 entrées qui rendent compte de 158 concepts (éléments et propriétés de l'application) et de 36 requêtes (demande de tarifs, d'horaires, etc.). Il faut comprendre que les dépendances décrites dans le lexique correspondent à des usages définis dans des requêtes standards. On peut regretter l'utilisation (implicite) de requêtes prédéfinies, qui obligent l'utilisateur à se conformer aux attentes du système. Toutefois, cette contrainte sera d'autant plus acceptable qu'on aura procédé à une analyse détaillée des usages langagiers avant de procéder à la conception du système. Nous verrons plus loin, avec le système LOGUS, que cette limitation peut être évitée sans perte d'efficacité.

La stratégie d'analyse repose sur l'algorithme défini par Sleator et Temperley. Eventuellement partielle, elle procède par expansion d'îlots en rattachant les éléments ayant des attentes compatibles.

Ce rattachement permet d'identifier les réparations qui n'ont pas encore été caractérisées. Un opérateur spécifique modélise le phénomène d'entassement paradigmatique qui résulte de ces réparations. Il permet le rattachement de plusieurs éléments selon la même relation de dépendances, que l'on soit en présence d'une réparation (reparandum et altération) ou d'une énumération.

En fin d'analyse, un système de coût permet de hiérarchiser les différents graphes de dépendances obtenus. Nous privilégions, par ordre de coût décroissant, les analyses complètes, les analyses partielles avec îlots complets puis les analyses partielles avec éléments isolés. La représentation sémantique finale (structure de liens) est obtenue par parcours du graphe de dépendances non orienté qui a été construit.

A titre d'exemple, la figure 4.5 donne la structure de dépendances de l'énoncé (4.3) telle que l'a construite ROMUS. Les dépendances internes aux chunks sont factorisées sur cette illustration. Par exemple la séquence de mots « *le tarif* » est identifiée à un groupe nominal déterminant (prise en compte du déterminant) de tête lexicale *tarif*. On remarque que la réparation présente en début d'énoncé a bien été identifiée.

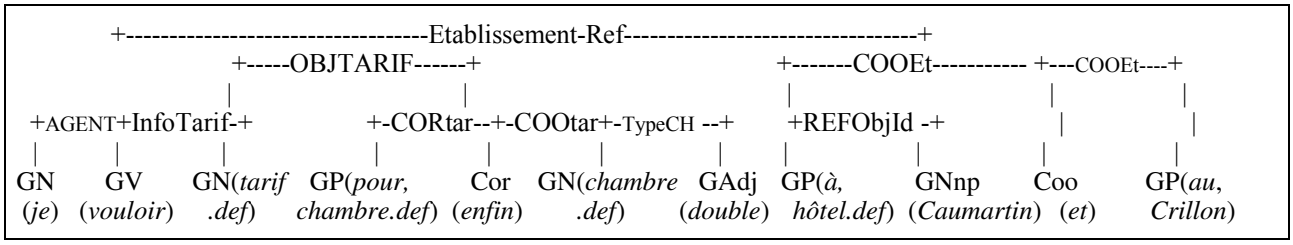

Figure 4.5 *— Exemple de structure de liens obtenue en sortie du système ROMUS*

1.5.5. Système LOGUS

Le système LOGUS[58] repose sur l'utilisation de techniques très originales dans le domaine de la communication parlée. Il s'agit en effet d'un système logique utilisant le λ-calcul pour construire une formule sémantique à la Montague[59]. Cette approche atypique répond à plusieurs justifications :

- Tout d'abord, nous avons vu que la dernière étape de traitement du système ROMUS utilisait le formalisme des grammaires de dépendances. Or, plusieurs travaux ont montré qu'il existait une équivalence formelle entre les grammaires de dépendances et des approches logiques telles que

[58] Villaneau J. (2003) Contribution au traitement syntaxico-pragmatique de la langue naturelle parlée: approche logique pour la compréhension de la parole. Doctorat l'Université de Bretagne Sud, Vannes, France. 6 décembre 2003. Rapport de recherche VALORIA-CORAIL-2003-02 ; Villaneau J., Antoine J.-Y., Ridoux O. (2002) LOGUS, un système formel de compréhension du français parlé spontané. Actes TALN'2002, Nancy, France, vol. 1, 165-174 ; Villaneau J., Antoine J.-Y., Ridoux O. (2001) Combining syntax et pragmatic knowledge for the understanding of spontaneous spoken utterances. Actes *4th International Conference on the Logical Aspects of Computational Linguistics*, *LACL'01,* Le Croisic, France. In *LNAI 2099*, Springer Verlag, 279-295 ;

[59] Montague R. (1974) Formal philosophy. Yale University Press, Newhaven ; Pour une présentation succinctes, on pourra consulter le chapitre consacré à la sémantique de Montague dans l'ouvrage de synthèse : Bouillon P. (1998) Traitement automatique des langues naturelles. Duculot, Bruxelles, Belgique. 128-131

les grammaires catégorielles[60].

- Ensuite, cette approche logique est compatible avec de nombreux travaux portant sur le dialogue. Je pense en particulier à la formalisation logique de la théorie des actes de langage proposée par Daniel Vanderveken (logique illocutoire[61]) ou encore à la *Discourse Representation Theory* (DRT) de Kamp et Reyle[62]. D'une manière générale, le recours à une modélisation logique du dialogue — le plus souvent dans le cadre de la logique des prédicats du 1er ordre — est fréquent en communication homme-machine[63]. LOGUS vise donc à terme une intégration forte de la compréhension et de la gestion du dialogue.

Quel que soit l'intérêt de cette modélisation, LOGUS répond aux caractéristiques suivantes.

Etiquetage — Cette étape se limite à un simple appel au dictionnaire de l'application. Elle consiste à associer à chaque mot sa ou ses définitions. Chaque mot est représenté par un triplet <*C,R,S*> où :

- *C* est la catégorie syntaxique du mot. Elle correspond à sa partie du discours, éventuellement enrichie par des propriétés morphosyntaxiques. Cette catégorie peut être simple (*nom_commun* par exemple) ou fractionnaire, au sens des grammaires catégorielles. Un déterminant (élément sous-catégorisé) est ainsi représenté par la catégorie fractionnaire *gn/nom_commun* qui signifie que son rattachement par la droite à nom commun permet de construire un groupe nominal. De même tout modifieur est décrit par une catégorie fractionnaire de type C/C — ou plus rarement C\C — où C est la catégorie modifiée[64].
- *R* définit le rôle du lexème dans l'univers de la tâche. Il peut être également être simple ou fractionnaire. (*objet)* ou encore *(prop quantité)*, pour un élément caractérisant une propriété de quantité, sont des exemples de rôles simples.
- *S* représente la sémantique (à la Montague) du mot considéré. Elle est décrite par un λ-terme.

La figure 4.6 donne quelques exemples de définitions sous forme de triplets <C,R,S>.

hôtel	<nom_commun, objet, hotel>
cher	<adj,(prop tarif), cher>
visiter	<infinitif,objet,visite>
pas	<adj/adj , (prop R)/(prop R), λ x.(pas x)>
la	<gn/nom_commun, R/R, λx.(obj x (deter def sing))>
un	<gn/nom_commun, R/R, λx.(obj x (deter indef sing))>

***Figure 4.6** — Dictionnaire du système ROMUS : quelques exemples de définitions*

Une même entrée lexicale peut correspondre à plusieurs triplets. Contrairement au système ROMUS, LOGUS n'opère aucune désambiguïsation à ce stade. Les ambiguïtés éventuelles sont levées par les étapes ultérieures d'analyse.

---

60 Lecomte A. (1996) Grammaire et théorie de la preuve : une introduction. *TAL*, 37(2). 1-38 ; König E. (1996) Introduction to categorial grammars. rapport de recherche, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart Allemagne.

61 Vanderveken D. (1994) A complete formulation of a simple logic of elementary illocutionary acts. In Tsohatzidis S. L. (Ed.) *Foundations of speech act theory : philosophical and linguistic perspectives*. Routldege. 99-131.

62 Kamp H., Reyle U. (1993), *From discourse to logic*. Kluwer Academic Publ, Amsterdam, Pays-Bas.

63 Pettier J.C., Guyomard M. (2000) Action modelling in dialogue context. Actes *3th International Workshop on Human-Computer Conversation*. Bellagio, Italie. 136-141 ; Villaseñor L., Caelen J. (1999) Une logique pour le dialogue coopératif homme-machine. Actes *2ème Colloque International sur l'Apprentissage Personne - Système*, *CAPS'98*, Caen, France. 53-62.

64 On retrouve une distinction assez proche entre éléments sous-catégorisés et modifieurs dans les grammaires d'arbres adjoints (arbres initiaux et arbres auxiliaires) : Abeillé A.(1993) Les nouvelles syntaxes : grammaires d'unification et analyse du français, Armand Colin, Paris

On remarque également qu'il existe un fort isomorphisme entre la catégorie syntaxique C associée à un mot et son rôle sémantique R. De fait, c'est cette connaissance syntaxique qui guide de facto la segmentation. Comme pour ROMUS, cette seconde étape est donc totalement générique.

Segmentation — La segmentation en *chunks* consiste à regrouper les mots de l'énoncé par composition de leurs λ-termes. Compte tenu du caractère local de l'analyse, cette composition ne requiert que l'utilisation des règles d'application droite et gauche des grammaires catégorielles de type AB[65] :

(A1) $A, A\backslash B \Rightarrow B$ (A2) $B/A, A \Rightarrow B$

Plus précisément, ces règles ont été adaptées pour pouvoir manipuler les triplets <R,S,T> :

(B1) $\langle C_1,R_1,S_1\rangle , \langle C_1\backslash C_2, R_1\backslash R_2, (\text{abst } F)\rangle \Rightarrow \langle C_2, R_2, (F\ S_1)\rangle$

(B2) $\langle C_1/ C_2,R_1/ R_2,(\text{abst } F)\rangle , \langle C_2, R_2, S_2)\rangle \Rightarrow \langle C_1, R_1, (F\ S_2)\rangle$

Ces règles restent parfaitement générales. Le fait qu'elles manipulent des informations sémantiques parallèlement aux catégories syntaxiques de la grammaire catégorielle n'a aucune influence sur la généricité de la segmentation.

L'exemple ci-dessous (figure 4.7) illustre la construction du groupe adjectival *pas trop cher*, qui nécessite un double appel à la règle d'application droite. Il montre comment la catégorie syntaxique, le rôle sémantique et la représentation sémantique du segment sont élaborés en parallèle. En particulier, on observe que la réduction des catégories fractionnaires syntaxiques (ici, adj / adj) et sémantiques (ici, prop R / prop R) se traduit systématiquement par l'application de l'abstraction correspondante (λx F) à la sémantique du terme atomique associé.

pas	<adj/adj ,	(prop R)/(prop R),	λx.(pas x)>
trop	<adj/adj,	(prop R)/(prop R),	λx.x>
cher	<adj,	(prop tarif),	cher>
pas trop cher	<adj,	(prop tarif),	(λx.(pas x) (λx.x cher)) $\equiv_\beta$(pas cher)>

***Figure 4.7** — Segmentation dans le système LOGUS : construction du triplet <C,R,S> représentant le segment adjectival « pas trop cher ». Le symbole $\equiv_\beta$ caractérise l'opération de β-réduction qui est directement implémenté dans λ-Prolog*

La stratégie d'analyse consiste à envisager toutes les compositions possibles. Elle est associée à une heuristique finale de filtrage qui privilégie les hypothèses comportant un nombre minimal de constituants.

On élimine les éléments qui sont isolés à l'issue de la segmentation, c'est-à-dire ceux qui correspondent à des catégories fractionnaires non réduites. Comme pour ROMUS, ces cas de rejet sont limités et correspondent à des amorces de réparations très fragmentaires. De même, la caractérisation complète des réparations est à la charge du niveau suivant de traitement.

Dépendances — Les possibilités de rattachement entre segments sont décrites par des prédicats logiques qui constituent une connaissance pragmatique spécifique à l'application. Un jeu de règles portant sur les triplets *<C, R, S>* permet de composer progressivement les différents segments pour construire la représentation sémantique de l'énoncé. Cette formule logique est donnée par l'élément S du triplet obtenu en fin de composition.

[65] Bar-Hillel Y. (1964). Language and information. Addison-Wesley, Reading, Etats-Unis ; Moorgat M. (1997) Categorial type logics. *In* van Benthem J., ter Meulen A. (Eds.) *Handbook of logic and language*. Elsevier Sciences, North-Holland, Amsterdam, Pays-Bas, 93-177.

Comme pour ROMUS, on définit des règles d'associations propres à la caractérisation des relations entre objets et des règles plus générales de coordination. Les réparations sont traitées de manière similaire. La règle (C) ci-dessous gère par exemple les répétitions avec enrichissement lexical du type « *un hôtel deux étoiles un hôtel pas trop cher quoi* » :

(C) $<C, objet, O_1>$, $<C, objet, O_2>$ $\Rightarrow$ $<C, objet, O_1 \cup O_2>$

id_etiquette(O_1,O_2)

On remarque que ce type de règle reste encore relativement générique.

Enfin, certains segments sont des marqueurs de l'acte de dialogue porté par l'énoncé. Leur analyse par un ensemble de règles spécifiques permet une première identification de cet acte.

La mise en œuvre des règles fait intervenir trois niveaux d'applications successifs suivant une stratégie d'analyse incrémentale propre au TAL robuste. Ces étapes se différencient par la granularité des éléments traités :

- la première étape correspond à l'identification des éléments de la tâche référencés par l'énoncé. Ces éléments peuvent correspondre à un segment unique, ou plus fréquemment à la combinaison de plusieurs chunks.
- la seconde étape permet l'identification des différentes propositions de l'énoncé. La notion de proposition doit être entendue d'un point de vue pragmatique. Il s'agit d'une partie de l'énoncé, groupée autour d'un noyau sémantique, qui porte un acte de dialogue. Seuls les énoncés multiples comportent donc plusieurs propositions.
- la dernière étape recherche finalement d'éventuelles dépendances entre propositions (coordination logique par exemple).

Chaque étape fait également l'objet d'une analyse incrémentale. Je vais me contenter de décrire ici la première étape d'identification des éléments de l'énoncé qui fait intervenir trois passes d'analyse :

- la première passe correspond à l'association de segments juxtaposés décrivant des objets simples de la tâche. Elle gère par exemple l'association du chunk nominal et du chunk adjectival dans la séquence [*un hôtel*] [*pas trop cher*]. Ces nouveaux segments combinés sont proposés au niveau suivant d'analyse.
- la seconde passe correspond à l'association des segments représentant des objets simples de la tâche pour décrire, si nécessaire, certains objets plus complexes. Ce niveau considère par exemple les chaînes d'objets telles que [*un hôtel*] [*près de la gare Montparnasse*] ou encore des objets coordonnés comme [*un hôtel*] [*ou*] [*une ferme auberge*].
- la troisième passe permet des compositions similaires au niveau précédent, avec pour seule différence une augmentation de la flexibilité d'association (variabilité dans l'ordre des segments, par exemple). Ce relâchement de contraintes est utilisé lorsque le système ne parvient pas à élaborer une représentation sémantique complète de l'énoncé.

La figure 4.8 donne à titre illustratif la formule logique qui correspond à la représentation sémantique de l'énoncé (4.3).

```
((requete vouloir)  (et  (de (tarif [ ])  (de  (chambre [(taille_chambre double)])
                                          (de  (hotel [identification 'Caumartin']))))
                         (de (tarif [ ])  (de (chambre [(taille_chambre double)])
                                              (de  (hotel [identification 'Crillon'])))  )  )
```

Figure 4.8 *— Exemple de formule logique obtenue en sortie du système LOGUS*

On relèvera pour conclure qu'il est possible d'aboutir à une analyse partielle en présence d'énoncés

fortement perturbés. Enfin, cette étape permet une résolution précoce des co-références anaphoriques par consultation de l'historique des énoncés utilisateurs.

1.5.6. Résultats expérimentaux

Comme l'ont montré ces deux présentations, les systèmes LOGUS et ROMUS présentent de fortes similitudes par delà leurs différences d'implémentation. Il n'est donc pas étonnant qu'ils présentent des résultats expérimentaux assez proches.

Tout d'abord, les systèmes ont participé à la campagne d'évaluation par défi du GDR-I3 où leurs comportements (points forts et faiblesses) furent assez proches. En particulier, ils ont présenté une robustesse appréciable sur les structures linguistiques complexes (à l'exclusion de certaines requêtes multiples) ainsi que sur des phénomènes de l'oral spontané tels que les répétitions, les corrections ou les dislocations (tableau 4.5 page suivante).

Tableau 4.5 — *Synthèse du comportement des systèmes ROMUS et LOGUS au cours de la campagne d'évaluation par défi du GDR-I3 (☺ : gestion satisfaisante des problèmes ; ☹ : gestion des problèmes perfectible).*

Système	ROMUS (version 2001)	LOGUS (version 2001)
erreurs de RAP	☹	☹
complexité	☺ sauf portée des négations sur les requêtes multiples	☺ sauf énoncés multiples de type requête + information
oral spontané	☺ sauf incises à l'intérieur d'un chunk	☺ sauf faux-départs
Dislocations	☺	☺

Ces systèmes ont connu depuis de nouvelles évolutions qui se traduisent par des améliorations significatives de leur comportement. En témoigne par exemple la comparaison des performances des deux dernières versions du système LOGUS sur les jeux de tests de la campagne d'évaluation par défi (tableau 4.6). A l'instar des évaluation de type ATIS (référence minimale et maximale), ce tableau distingue deux cas de compréhension correcte :

- *compréhension complète* — la structure sémantique de l'énoncé a été totalement extraite,
- *compréhension incomplète* — certaines éléments de l'énoncé, jugés non essentiels, n'ont pas été identifiés.

Tableau 4.6 — *Comparaison des performances des deux dernières versions du système LOGUS sur les jeux de tests de la campagne d'évaluation par défi du GDR-I3 : taux d'énoncés correctement compris par séries de test. Chaque système est évalué sur un jeu de tests spécifique.*

Type de difficulté	ROMUS (version 2003)	LOGUS (version 2001)	LOGUS (version 2003)
Réparations, dislocations	**97,9 %** (dont 4,4 % incomplets)	**96,1 %** (dont 3,5 % incomplets)	**98,0 %** (dont 4,3 % incomplets)
Enoncés complexes	**96,9 %** (dont 2,0 % incomplets)	**68,3 %** (dont 16 % incomplets)	**89,7 %** (dont 16,7 % incomplets)
Phénomènes multiples	**79,2 %** (dont 18,8 % incomplets)	**64,3 %** (dont 16,7 % incomplets)	**85,4 %** (dont 23,7 % incomplets)

Les taux de robustesse fournis regroupent ces deux situations. Suivant les principes d'évaluation discriminante de la campagne DEFI, ces résultats sont regroupés par séries de tests. La première série regroupe des énoncés présentant des procédés de l'oral spontané ou des dislocations tandis que la seconde concerne des énoncés structurellement complexes (requêtes multiples, rattachement

d'arguments récursifs, etc.). La dernière série, qui regroupe ces deux types de difficultés dans chaque test, concerne des énoncés qui se rapprochent de l'interaction orale entre interlocuteurs humains.

Les performances plus mitigées que l'on observe sur cette troisième série sont à la mesure de la difficulté que représente le traitement d'énoncés libres proches de la conversation humaine. D'une manière générale, les résultats obtenus par les deux systèmes sont toutefois très encourageants et soutiennent la comparaison avec les autres participants[66]. Ils semblent confirmer la pertinence des stratégies d'analyse adoptées par ces systèmes.

Par ailleurs, nous avons cherché à évaluer en interne l'influence des erreurs de reconnaissance sur le comportement du système ROMUS. Partant de 600 énoncés oraux traités par une dictée vocale grand public (IBM Via Voice), nous avons obtenu un corpus de test comportant :

- 358 énoncés présentant une ou plusieurs erreurs de reconnaissance de type lexical (insertion, substitution ou élision)
- 140 énoncés présentant au moins une erreur d'accord.

Ces erreurs ont eu une influence sur la compréhension dans 16% des cas. Elles n'ont cependant jamais conduit à une représentation erronée de l'énoncé, mais simplement à la construction de structures incomplètes. Cette expérimentation a bien entendu été réalisée dans un cadre réducteur. J'espère toutefois qu'elle est relativement représentative du comportement du système en sortie de reconnaissance. Nous envisageons de réaliser une expérimentation analogue avec le système LOGUS. Cette évaluation serait conduite sur un système de reconnaissance de recherche, dans le cadre d'une collaboration avec le laboratoire IRISA/Cordial (Lannion).

1.5.7. Conclusions

En dépit de leur caractère limité, ces résultats expérimentaux sont encourageants. En particulier, la campagne d'évaluation par défi du GDR I3 a montré qu'il est possible de viser une compréhension détaillée de la parole sans perdre en robustesse d'analyse. De ce point de vue, les systèmes LOGUS et ROMUS présentent des performances équivalentes aux approches sélectives. Ces résultats sont donc une incitation à poursuivre dans la voie d'une compréhension à fort ancrage linguistique.

Comme je l'ai expliqué plus haut (cf. § 1.3), cette approche est motivée par la recherche d'une généralisation du dialogue oral homme-machine à des contextes applicatifs plus riches. La campagne d'évaluation par défi du GDR I3 ne nous a pas permis d'étudier cette question, puisque les systèmes sélectifs en compétition n'ont pas été testés sur des tâches complexes[67]. A l'opposé, la prochaine campagne d'évaluation MEDIA (projet TECHNOLANGUE) portera sur le renseignement touristique. Les systèmes LOGUS et ROMUS participeront à ce projet qui concernera en outre la compréhension de la parole en contexte. Réunissant l'ensemble des laboratoires français travaillant sur cette thématique, il devait permettre de mieux situer l'apport de nos travaux par rapport aux méthodes sélectives.

Pour l'heure, j'espère que cette présentation aura montré la cohérence scientifique de nos travaux. Cette démarche scientifique se retrouve dans nos recherches portant sur l'aide au handicap.

2. SYSTEMES D'AIDE A LA COMMUNICATION LANGAGIERE

J'ai orienté depuis trois ans une partie des mes activités de recherche sur l'aide à la communication pour personnes handicapées. Cette problématique intéressante est souvent méconnue des chercheurs en ingénierie des langues. Il me paraît donc important de la situer dans un paragraphe introductif.

---

[66] Rappelons que la méthodologie d'évaluation par défi ne permet pas de comparaison directe des différents systèmes (chapitre 3, § 2.4). ROMUS et LOGUS ont eux-mêmes été évalués sur des jeux de tests différents.

[67] Les tâches étudiées par ces systèmes étaient respectivement le renseignement d'horaires de train, le renseignement ferroviaire (horaires et réservation) et la réservation hôtelière (cf . chapitre 3, tableau 3.4). Comme je l'ai montré au chapitre 2 (tableau 2.3), ces tâches se traduisent par une complexité très sensiblement inférieure à celle du renseignement touristique.

2.1. Handicap et ingénierie des langues[68]

Contrairement aux autres champs d'application de l'ingénierie des langues, l'aide au handicap ne se développe que marginalement alors qu'elle répond à une attente sociétale forte. On constate ainsi que moins d'une dizaine d'équipes francophones, souvent de tailles très réduites, développent des travaux sur le handicap là où le réseau francophone des industries de la langue (FRANCIL) regroupe plus de 60 centres de recherche. Ce déséquilibre se retrouve du côté des applications industrielles, qui sont le plus souvent développées par des PME aux ressources humaines et financières limitées. Les aides logicielles à la communication disponibles sur le marché présentent ainsi des fonctionnalités largement inférieures à celles de systèmes de recherche, alors que nombre d'entre eux sont opérationnels et ont fait l'objet d'expérimentations auprès de patients handicapés.

Cette situation est d'autant plus regrettable que l'enjeu économique de l'aide au handicap est loin d'être négligeable. Près d'un français sur douze est ainsi une personne handicapée et plus de trois millions de français bénéficient actuellement d'une aide humaine ou technique au handicap. On sait par ailleurs que le vieillissement de la population observé dans les pays développés se traduit par une forte augmentation du nombre de personnes âgées dépendantes. Une des actions clés du 5ème PRCD de la Commission Européenne concerne précisément la thématique « vieillissement de la population et handicaps ». 2003 a par ailleurs été déclarée année européenne du handicap.

L'aide au handicap représente par ailleurs un enjeu social considérable, comme en témoignent les attentes très fortes des personnes handicapées en matière d'assistance technique. Compte tenu de l'importance de la communication langagière dans les sociétés humaines, l'ingénierie des langues est un des domaines technologiques les plus interpellés par ce problème. Le développement d'Internet, véritable portail vers le monde extérieur pour les personnes isolées et dépendantes, ne fait qu'accroître cette demande sociale.

Afin d'illustrer la diversité des attentes des personnes handicapées, citons quelques exemples d'applications potentielles de l'ingénierie des langues dans le monde du handicap :

- **communication assistée par ordinateur**, sur laquelle nous reviendrons par la suite.
- **transfert de modalité d'entrée** (compensation en temps réel de la modalité défaillante) : synthèse de parole à partir de texte[69], sous-titrage par reconnaissance vocale, …
- **assistance à base de langages spécialisés** : langages de commande pour la robotique d'assistance[70], aide à la programmation informatique, ...

L'intérêt scientifique de l'aide au handicap est par ailleurs indéniable. Il s'agit d'une problématique exemplaire où la prise en compte de l'utilisateur est essentielle. Des notions telles que la robustesse d'analyse, la personnalisation et l'adaptation, l'évaluation et la validation ergonomique des systèmes, revêtent ici une dimension incontournable. On conçoit donc l'importance de cette problématique dans la perspective d'une ingénierie des langues qui replace l'utilisateur au centre de ses préoccupations. L'aide au handicap est ainsi susceptible de nourrir l'ensemble des recherches

[68] Ce paragraphe est extrait, sous forme augmentée, du texte de présentation de l'atelier *Ingénierie des Langues et Handic*ap : Antoine J.Y., Le Pévédic B. (2001) Ingénierie des langues et handicap. Actes *TALN'2001*, conférence associée *Ingénierie des Langues et Handicap*, Tours, France. vol. 2, 179-182.

[69] Ricco X., Dutoit T. (2001) Vers un logiciel multilingue et gratuit pour l'aide aux personnes handicapées de la parole : HOOK (une interface du projet W). Actes *TALN'2001*, conférence associée *Ingénierie des Langues et Handicap*, Tours, France. vol. 2, 223-232.

[70] Richard P., Gaucher P., Maurel D. (2000), Projet CNHL : Chambre Nomade pour Handicapés Lourds, Actes *Handicap'2000*, Paris, France, pp. 101-107.

menées en ingénierie des langues[71]. Je reviendrai sur ce point en conclusion (cf. § 2.4) de mes travaux dans le domaine de la communication langagière assistée par ordinateur.

2.2. Communication assistée par ordinateur : problématique[72]

L'aide à la communication langagière, encore appelée communication assistée par ordinateur, s'adresse à des personnes dont l'usage de la parole est fortement ou complètement altéré et dont les facultés motrices sont très réduites (tétraplégie). Outre les nombreux accidentés de la route, c'est le cas des patients IMC (Infirmes Moteurs Cérébraux), SLA (Sclérose Latérale Amyotrophique) ou présentant un *Locked-In Syndrome*. Dans tous les cas, le handicap physique est très sévère, la communication est privée de son support oral habituel et les modalités d'interaction sont limitées.

Un moyen de communication alternatif est le recours aux systèmes dits de communication assistée (AAC ou *Alternative and Augmentative Communication* en anglais). Il s'agit de systèmes de suppléance dont le rôle est d'augmenter ou de restaurer, ne serait-ce que partiellement, la fonction de communication. Deux types de communication sont concernés par ces systèmes :

- **la communication d'urgence ou de nécessité**, qui se limite à l'expression de besoins vitaux de la vie quotidienne. Dans ce cas, l'aide vise essentiellement une plus grande autonomie de la personne handicapée. Elle est guidée par une exigence de rapidité et d'efficacité (message clair et facilement compréhensible).
- **la communication langagière** en général, qui regroupe toutes nos interactions écrites et orales avec nos semblables. Ici, l'aide vise une amélioration de la qualité de vie de la personne handicapée, en favorisant son intégration dans la société. Son objectif est de rapprocher les capacités interactionnelles des handicapés de celles des personnes valides. Elle est donc guidée par une exigence de richesse et de naturalité des productions langagières.

Si l'accès à l'autonomie est une nécessité primordiale pour les personnes handicapées, celles-ci sont de plus en plus demandeuses de systèmes de suppléance qui leur permettent de mener une vie plus « ordinaire ». L'aide à la communication langagière générale est donc appelée à répondre à des besoins sociaux de plus en plus marqués.

Comme je l'ai dit plus haut, les utilisateurs de ces systèmes de suppléance présentent des capacités d'interaction avec le monde physique très limitées. Le plus souvent, celles-ci se limitent à un geste à un seul degré de liberté (souffle, clin d'œil, inclinaison de la tête, mouvement de la main) qui est par ailleurs assez mal contrôlé. Dans ces situations difficiles, la communication — qu'elle soit de nécessité ou langagière — repose sur l'écriture de phrases à l'aide d'un tableau de symboles. Le message est construit en sélectionnant successivement les différents symboles qui le composent. L'intervention de la personne handicapée se limite à la désignation des symboles, l'ordinateur étant chargé de relever les éléments désignés. Les symboles sont souvent les lettres de l'alphabet orthographique (éventuellement réduit). Cet usage n'est cependant pas exclusif. Dans le cas de

[71] J'observe à ce sujet que l'ensemble des techniques développées par l'ingénierie des langues (modèles de langage probabilistes, systèmes à base de connaissances, grammaires probabilisés, etc.) a déjà fait l'objet d'une utilisation dans le monde du handicap. Comme le notent Philippe Boissière et Daniel Dours « *il faut très certainement voir [dans cette diversité] le fait que chaque personne handicapée [est] un cas particulier nécessitant presque un système spécifique* » : Boissière P., Dours D. (2001) Comment VITIPI, un système d'assistance à l'écriture pour les personnes handicapées peut offrir des propriétés intéressantes pour le TALN. Actes *TALN'2001*, conférence associée *Ingénierie des langues et handicap*, Tours, France. vol. 2, 183-192.
Cette hétérogénéité est certainement un frein au développement des systèmes industriels destinés aux handicapés. D'un point de vue scientifique, elle fait au contraire du monde du handicap un champ de validation et de comparaison par excellence pour l'ingénierie des langues

[72] Ce paragraphe est extrait, sous forme étendue: Schadle I., Antoine J-Y., Le Pévédic B., Poirier F., (2002) SibyLettre, prédiction de lettres pour la communication augmentée, revue *RIHM*, Vol3, n°2, Europia, Paris, France, 2002.

personnes ne disposant pas (ou plus) des capacités cognitives nécessaires à la maîtrise de la langue, on a recours à un alphabet phonétique voire à des pictogrammes[73] ou des symboles BLISS[74].

Historiquement, l'aide à la communication a tout d'abord reposé sur la construction de systèmes physiques spécifiques (tableau électronique avec synthèse de parole intégrée par exemple[75]). Elle est désormais envisagée sous la forme de systèmes logiciels installés sur l'ordinateur personnel de la personne handicapée. On distingue alors l'interface matérielle d'accès à l'ordinateur, de la partie logicielle proprement dite (figure 4.9).

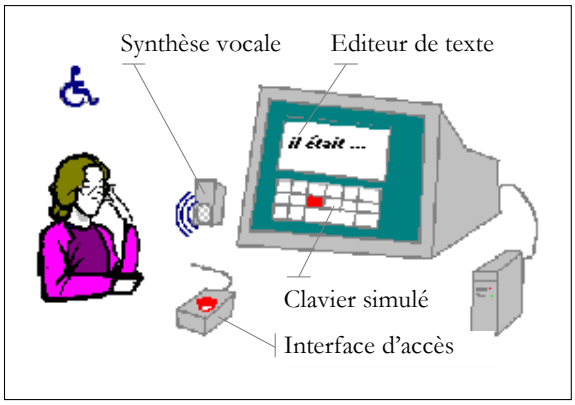

Figure 4.9 *— Système d'aide à la communication assisté par ordinateur.*

L'interface matérielle dépend du geste libre laissé par le handicap. Elle remplace le clavier réel, interface rendue inadaptée. Il peut s'agir d'un joystick, d'une commande oculaire, d'une commande par souffle, d'un simple bouton poussoir, etc. Comme nous l'avons dit, les personnes souffrant d'un handicap sévère ne peuvent souvent réaliser qu'un geste mal contrôlé à un seul degré de liberté. L'interaction avec l'interface matérielle se limite alors à une pression « tout ou rien » qui ne laisse même pas la possibilité d'une discrimination entre clic bref et clic long.

La partie logicielle constitue le système d'aide à la communication proprement dit. Son interface est constituée principalement d'un clavier simulé (présenté à l'écran) et d'un traitement de texte, ou d'une simple zone d'édition qui affiche le texte en cours de saisie. Dans le cas de la communication orale, une synthèse vocale permet de prononcer le texte composé[76].

La saisie sur clavier simulé est étroitement liée au nombre de degrés de liberté offert par l'interface matérielle. Dans le cas d'un handicap n'autorisant que des commandes « tout ou rien », la sélection des symboles ne peut être réalisée que par un défilement automatique du curseur qui met en surbrillance les touches une à une. Lorsque le curseur désigne la lettre désirée, l'utilisateur la valide via l'interface matérielle. Ce mode de saisie est bien entendu excessivement lent.

Tableau 4.7 *— Exemples de vitesses de communication (mots/minute)* [77]

---

73 Brangier E., Gronier G. (2000) Conception d'un langage iconique pour grands handicapés moteurs aphasiques. Actes *Handicap'2000*. Paris, France. 93-100 ; Abraham M. (2000) Reconstruction de phrases oralisées à partir d'une écriture pictographique. Actes *Handicap'2000*, Paris, France. 151-156 ; Vaillant, P. (1997) PVI : Système de traduction d'icônes en langue, interaction entre modalités sémiotiques. Doctorat U. Paris-XI, Orsay, France.

74 Toulotte J.M., Baudel-Cantegrit B., Trehou G (1990) Acceleration method using a dictionary access in a BLISS communicator. Actes *1990 Biennial Conference of the International Society for Augmentative and Alternative Communication*. Stockholm, Suède.

75 Voir par exemple le système SPARTE mis au point au CNET Lannion dans les années 80 : Glandière M. (1983) SPARTE. *Education et informatique*, 17.

76 Voir par exemple le système HOOK qui bénéficie de l'expérience de l'Université de Mons en matière de synthèse de parole à partir du texte : Ricco X., Dutoit T. (2001) Vers un logiciel multilingue et gratuit pour l'aide aux personnes handicapées de la parole : HOOK (une interface du projet W). Actes *TALN'2001*, conférence associée *Ingénierie des Langues et Handicap*, Tours, France. vol. 2, 223-232

77 Données extraites de : Vincent - Le Pévédic B. (1997) Prédiction morphosyntaxique évolutive dans un système d'aide à la saisie de textes pour des personnes handicapées physiques : HandiAS. Thèse de doctorat de l'Université de Nantes, Nantes, France. 13 octobre 1997.

Type de communication	Vitesse de saisie (mots/minute)
Communication orale	150
Communication manuscrite	12-16
Saisie secrétaire	25
Saisie clavier un doigt	11
Handicapé avec aide purement mécanique	5

Le tableau 4.7 présente ainsi quelques vitesses de communication relevées dans la littérature. On constate que la vitesse d'un handicapé qui ne dispose que d'un système d'aide mécanique (i.e. accès au clavier simulé sans traitements linguistiques complémentaires) est au mieux 30 fois inférieure à celle d'une interaction orale entre personnes valides.

On mesure dès lors l'intérêt des systèmes d'aide à la communication intégrant des traitements linguistiques pour accélérer la saisie. Cet apport ne se limite pas aux gains en temps de saisie. Pour un handicapé, la saisie de texte est une activité fastidieuse qui requiert une forte charge cognitive. Il en découle la production de nombreuses fautes et une fatigue importante des patients. Les systèmes de suppléance doivent donc permettre une amélioration qualitative et quantitative des capacités de communication des handicapés.

2.3. Traitements linguistiques pour une aide logicielle à la communication

Deux approches complémentaires sont envisageables pour accélérer la saisie. La première cherche à minimiser le nombre de saisies, la seconde à accélérer la sélection sur le clavier simulé.

2.3.1. Réduire le nombre de saisies

Différentes approches peuvent être envisagées pour économiser le nombre de saisies.

- **Rappel de phrases préenregistrées par touches de fonction** — Cette aide efficace mais limitée ne fait appel à aucun traitement linguistique. Elle est destinée à la communication d'urgence et de nécessité (phrases types). Elle peut être intégrée à n'importe quel système pour permettre la composition rapide de messages au caractère urgent irrépressible.
- **Systèmes de désabréviation** — Cette approche, que l'on retrouve en sténographie, consiste à limiter le nombre de saisies par l'utilisation d'un système d'écriture abrégée. La correspondance entre les mots et les abréviations peut être biunivoque, auquel cas la désabréviation se limite à un accès au lexique abrégé. Dans le cas contraire, un traitement linguistique contextuel est requis[78]. A ma connaissance, seuls des systèmes d'écriture non ambigus ont été utilisés à ce jour dans le domaine du handicap[79]. De nombreux aveugles utilisent au quotidien un système Braille abrégé. La maîtrise d'un système d'écriture abrégé requiert toutefois des capacités cognitives dont ne disposent plus la plupart des handicapés intéressés par les systèmes d'aide à la communication.

De fait, seule l'analyse du handicap et des capacités cognitives et motrices de la personne handicapée peut permettre de faire un choix entre une saisie par désabréviation et les méthodes prédictives décrites ci-dessous. D'une manière globale, les méthodes prédictives toucheront toutefois une population nettement plus nombreuse.

2.3.2. Accélérer la sélection par prédiction linguistique

Ici, l'utilisateur doit a priori saisir la totalité du message. L'idée est d'accélérer la saisie en prédisant les symboles à venir (lettres, lemmes ou mots entiers) les plus probables en fonction du contexte. Deux approches peuvent être envisagées pour réaliser cette prédiction :

[78] Les abréviations peuvent également être crées à la volée : McCoy, K. F., Demasco, P. (1995) Somme applications of natural language processing to the field of augmentative and alternative communication. Actes *IJCAI'95 Workshop on Developing AI Applications for Disabled People*, Montréal, Canada. 97-112.

[79] Voir par exemple le système HOOK (Ricco X., Dutoit T., 2001, *op. cit.*) qui repose sur l'abrégé Braille II. Certains systèmes diffusés librement (ShortHand, WiVik, etc.) utilise également des techniques de désabréviation.

- **Prédiction descendante à partir de schémas prédéfinis** — Cette approche propose de guider la communication à partir de scénarios prédéfinis. L'utilisateur se contente de saisir certains mots clés qui remplissent des trous dans les énoncés correspondant aux scénarios choisis. On parle de composition de phrases assistée par ordinateur[80]. Les travaux les plus représentatifs de cette démarche ont conduit à la réalisation du système ILLICO dans le cadre du projet européen KOMBE[81]. Dans ce système, l'aide à la composition repose sur une stratégie de génération descendante qui définit dynamiquement les contraintes lexicales, syntaxiques et sémantico-pragmatiques qui président à la génération de l'énoncé. Elle repose sur des grammaires qui rendent compte de schémas spécifiques à un thème donné.

  Les techniques de prédiction descendante permettent une génération efficace d'énoncés relativement riches. Elles présentent cependant deux limitations importantes à mes yeux. D'une part, on est en présence d'une prédiction finalisée, au sens où la génération se limite à des thèmes bien précis. Dans le cas de KOMBE, ces thèmes étaient orientés vers la relation patient / médecin. Il est bien entendu possible de multiplier les grammaires de génération sur différents domaines. Cette approche ne saurait cependant être généralisée à la conversation générale. On retrouve ici les problèmes de généricité rencontrées en compréhension de la parole finalisée.

  D'autre part, on peut estimer que la définition de scénarios communicationnels contraint fortement l'expression des utilisateurs. Or, les handicapés sont sensibles à cette restriction, ce qui expliquent qu'ils soient réticents à utiliser de ce type d'aide à la communication. Les techniques de prédiction descendante recoupent ici, de manière plus flagrante, les limitations des systèmes de compréhension reposant sur des schémas prédéfinis.

  En conclusion, il me semble que cette approche ne peut être envisagée que pour des situations spécifiques qui nécessitent une forte efficacité communicationnelle tout en mettant en jeu une interaction plus riche que celle de la communication d'urgence ou de nécessité. Le choix de la relation patient / médecin retenu dans le projet KOMBE m'apparaît de ce point de vue très pertinent. Il n'en reste pas moins que les besoins des handicapés ne se limitent pas à ces interactions très particulières. Je remarque d'ailleurs que les chercheurs du LIM, qui ont développé le système ILLICO, se tournent désormais vers des approches co-génératives. D'où l'intérêt d'une prédiction ascendante portant sur la langue générale.

- **Prédiction ascendante** — Cette approche est celle qui réunit actuellement le plus de chercheurs. Son objectif est de réaliser une prédiction dynamique qui s'appuie sur le contexte constitué par l'énoncé déjà saisi. La prédiction peut porter aussi bien sur la lettre que le mot à venir.

  De nombreux systèmes relèvent de cette approche[82]. Ils diffèrent à la fois par le type d'analyse linguistique sous-jacente à la prédiction, mais aussi par des considérations ergonomiques. En particulier, certains systèmes effectuent une complétion automatique des mots dès qu'il n'y a plus d'ambiguïté, tandis que d'autres proposent une liste de prédiction au sein de laquelle l'utilisateur peut piocher plutôt que de continuer la saisie. La première solution est mise à défaut dans le cas des mots hors vocabulaire, tandis que la seconde nécessite des opérations de sélection supplémentaires pour accéder à l'information recherchée.

---

80 Richardet N. (1998) Composition de phrase***s*** assistée - Un système d'aide à la communication pour handicapés. Thèse de doctorat, Université de la Méditerranée, Marseille, France.

81 Guenthner F., Krüger-Thielmann K., Pasero R., Sabatier P. (1992) Communication aids for ALS patients. Actes *3rd International Conference on Computers for the handicapped persons*, *ICCHP'1992*, Vienne, Autriche. 303-307 ; Guenthner F., Langer S., Krüger-Thielmann K., Pasero R., Richardet N., Sabatier P. (1992) KOMBE : communication aids for the handicaped. rapport technique 92-55, CIS, Universität München, Munich, Allemagne.

82 Boissière P., Dours D. (2001) Comment VITIPI, un système d'assistance à l'écriture pour les personnes handicapées peut offrir des propriétés intéressantes pour le TALN. Actes *TALN'2001*, conférence associée *Ingénierie des langues et handicap*, Tours, France. vol. 2, 183-192 ; Maurel, D. Rossi, N. Thibault, R. (2001) Handias : un système multilingue pour l'aide à la communication de personnes handicapées. Actes *TALN'2001*, conférence associée *Ingénierie des langues et handicap*, Tours, France. vol. 2, 203–212 ; Vincent - Le Pévédic B. (1997) op. cit. ; Magnuson, T. (1995) Word Prediction as Linguistic Support for Individuals with Reading an Writing Difficulties. Actes *TIDE : The European context for assistive technology*. Paris, France. 316-319.

Dans tous les cas de figure, l'économie de saisie reste limitée. Les systèmes VITIPI et HandiAS[83] annoncent par exemple des gains de saisie compris respectivement entre 26% et 43%. Ceci laisse encore plus de la moitié des lettres à saisir par l'utilisateur dans des conditions difficiles.

Mes travaux visent précisément à gagner en pouvoir de prédiction par une analyse linguistique plus fine des énoncés saisis. Mon souci est de permettre une communication la moins contrainte possible du point de vue de l'expression langagière et du thème abordé. C'est pourquoi j'ai orienté cette recherche vers la prédiction ascendante de mots.

La prédiction ascendante de mots requiert une prise en compte l'énoncé déjà saisi. A l'heure actuelle, cette analyse contextuelle est relativement restreinte. On en reste en effet à des traitements très superficiels qui font par exemple appel à des modèles de langages markoviens (*N-grams*). Le comportement très perfectible des applications actuelles me laisse penser qu'une prédiction fiable ne pourra être atteinte que par des systèmes dont les compétences linguistiques ne sont pas réduites à la portion congrue. C'est dans cet esprit qu'a été réalisé le système SIBYLLE d'aide à la communication langagière.

2.4. Système SIBYLLE

Le système SIBYLLE de communication assistée a été réalisé dans le cadre du doctorat d'Igor Schadle sous ma direction et celle de Franck Poirier[84]. La conception du système s'est effectuée en collaboration avec le Centre Mutualiste de Rééducation et Réadaptation Fonctionnelle (CMRRF) de Kerpape. Le centre de Kerpape apporte sa compétence dans le domaine du handicap et son expérience dans les systèmes de communication assistée. Son rôle est également de servir de centre de test et de validation des différents outils qui seront issus du projet. Dans le cadre de cette collaboration, de nombreux patients sont des enfants IMC, aux facultés motrices très réduites. Ils utilisent comme interface matérielle le bouton poussoir suivant une modalité « tout ou rien ».

2.4.1. Architecture générale et principes d'utilisation

Le système d'aide à la communication SIBYLLE[85] offre deux niveaux de prédiction présentés sur la figure 4.5.

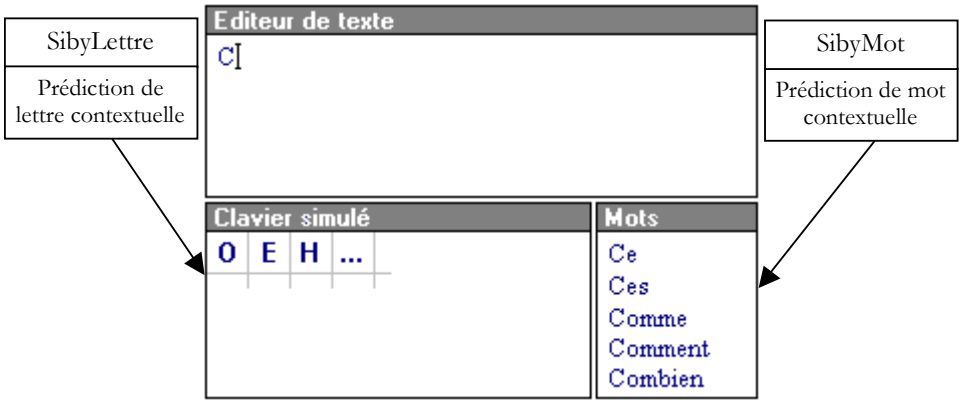

Figure 4.5 *— Principe de l'interface à clavier simulé du système Sibylle*

- une prédiction de mots (module *SibyMot*) qui s'appuie sur un modèle linguistique qui prend en considération le contexte de saisie. Elle fournit à l'utilisateur la liste des N meilleurs hypothèses lexicales. Cette liste est actualisée après chaque saisie de lettre. La prédiction repose sur une

[83] Boissière P., Dours D. (2001) *op. cit.* ; Vincent - Le Pévédic B. (1997) *op. cit* ;

[84] Brigitte Le Pévédic a également suivi le travail de thèse d'Igor.

[85] Pour une présentation plus détaillée du fonctionnement du système, avec des exemples de session d'utilisation, on consultera : Schadle I., Antoine J-Y., Le Pévédic B., Poirier F. (2002) SibyLettre, prédiction de lettres pour la communication augmentée, revue *RIHM*, Vol3, n°2, ISSN 1289-2963, Europia, Paris, France, 2002.

segmentation de l'énoncé en constituants minimaux non récursifs (*chunks*). Elle sera décrite au paragraphe 2.4.3.

- une prédiction de lettres (module *SibyLettre*) qui tient uniquement compte du mot en cours de saisie. Cette fonctionnalité est utile au début de la saisie des mots (deux premières lettres) lorsque le nombre de prédictions fournies par *SibyMot* reste trop importante. La prédiction de lettres est également utile sur les cas atypiques pour lesquels la prédiction lexicale est mise en défaut. Cette situation concerne en particulier les mots hors vocabulaire. C'est pourquoi la prédiction de lettres ne repose pas sur l'utilisation d'un lexique comme dans d'autres systèmes. Le module *SibyLettre* sera décrit au paragraphe 2.4.2.

Afin de limiter les opérations nécessaires à la sélection des symboles, la prédiction de lettres n'est pas affichée sous forme de liste, mais est intégrée directement dans le clavier simulé (figure 4.5). L'ordre de présentation des lettres sur le clavier est recalculé dynamiquement après chaque saisie. La prédiction de lettres est toujours réalisée en parallèle avec la prédiction de mots. Plusieurs expérimentations ont montré que les handicapés s'adaptaient sans problème au caractère dynamique de l'interface. Un seul patient, qui présentait des déficits de perception visuelle, n'a pas pu s'adapter au système.

L'interface du système SIBYLLE est paramétrable afin de s'adapter aux besoins de chaque handicapé. En particulier, la vitesse de défilement du curseur est modifiable, de même que la longueur de la liste prédiction (1 à 10 mots) fournie par *SibyMot*. Le passage du clavier simulé à la liste de prédiction lexicale s'effectue par sélection d'un symbole particulier intégré au clavier. La position de cette case est également paramétrable. Je vais maintenant présenter en détail ces deux modules de prédiction.

2.4.2. SibyLettre

Le module de prédiction de lettres est caractéristique de la démarche pragmatique auxquelles doivent recourir les recherches sur le handicap. L'objectif du module *SibyLettre* est, rappelons-le, de réaliser une prédiction ne faisant appel à aucune connaissance lexicale. Aussi ma première idée fût-elle de baser la prédiction sur une connaissance purement morphologique. Compte tenu de la taille limitée du contexte d'analyse (deux syllabes au maximum), j'ai également proposé d'étudier l'emploi d'un modèle markovien comme solution de contrôle. Les performances de ce système stochastique s'étant avérées excellentes, nous n'avons finalement pas développé de solution alternative. Cette expérience montre, comme je l'avais noté dans le chapitre premier, qu'une approche probabiliste peut s'avérer pertinente du moment que la tâche étudiée s'y prête.

La prédiction est basée sur l'estimation de la probabilité d'occurrence $Pr(L_i \mid L_{i-1}, \ldots L_{i-N+1})$ d'une lettre L_i donnée connaissant les N-1 lettres déjà saisies (modèle N-gram). Pour un contexte donné, les lettres sont classées par probabilités décroissantes sur le clavier simulé. Ces probabilités ont été estimées à partir de l'observation de cinq années du journal *Le Monde* (1995 à 1999). En cas d'absence d'observation d'un N-uplet, la probabilité résultante est calculée par repliement sur les probabilités d'ordre inférieur. Le caractère *espace*, qui joue le rôle de limite du mot, est un symbole du modèle. Notons à ce sujet que le contexte de prédiction est limité par le début du mot. C'est-à-dire que pour les P premières lettres du mot, avec $P \leq N$, la fenêtre de contexte est de taille P-1.

Tableau 4.8 *— SibyLettre : rang de prédiction moyen en fonction de la taille N du modèle*

N	**1**	**2**	**3**	**4**	**5**
rang moyen	7,1	4,7	3,8	3,2	2,9

SibyLettre a été soumis à la fois à une validation objective et subjective. L'évaluation quantitative a été menée sur le corpus du *Monde* suivant une méthodologie de validation croisée par années. Elle a consisté à calculer le rang moyen de la lettre attendue dans la liste ordonnée d'hypothèses établie par *SibyLettre*. Cette étude (tableau 4.8) montre que pour une fenêtre de contexte de 4 lettres en arrière (N=5), la lettre attendue est en moyenne à la troisième position. Pour un contexte d'une lettre

seulement (N=2), donc en début de saisie de mot, la lettre attendue est déjà dans les 5 premières propositions en moyenne. Lorsque l'on sait que l'alphabet retenu comportait 65 lettres (diacritiques et lettres d'autres alphabets latins), on mesure le gain de temps de saisie qu'autorise la prédiction.

Au final, le nombre de défilements pour atteindre une lettre avec *SibyLettre* est en moyenne de 2,9 pas (N=5). A titre de comparaison, ce nombre est de 7,1 pour un défilement linaire avec arrangement statique des lettres suivant leur fréquence dans la langue (cas N=1 du tableau 4.8). On peut montrer que le nombre de défilements est de 4,3 pour un défilement ligne/colonne qui nécessite également un clavier statique. Ces deux modes de saisie sont utilisés dans les systèmes d'aide à la communication commercialisés. On saisit donc l'amélioration que représente ce simple module de prédiction stochastique.

Cette amélioration se ressent fortement à l'usage. Elle a été confirmée au cours d'une évaluation qualitative qui a été menée au CRMMF de Kerpape et qui nous a permis de valider le principe d'un réarrangement dynamique des lettres sur le clavier simulé. Le gain apporté par *SibyLettre* est essentiellement apprécié en terme de confort d'utilisation par les handicapés. Le défilement linéaire et son unique validation ont été particulièrement appréciés par les patients, qui étaient habitués jusqu'ici à des claviers simulés statiques avec défilement ligne / colonne (double validation). Les enseignants de l'école primaire qui est intégrée au centre ont également constaté que les enfants qui utilisent *Sibylettres* composent plus de textes (rapidité de saisie accrue, mais aussi fatigue moins élevée). On observe également qu'ils commettent moins de fautes d'orthographe, la prédiction semblant orienter leurs productions vers des solutions correctes.

2.4.3. SibyMot

Si ces premiers résultats sont encourageants, c'est néanmoins le module de prédiction de mots qui fait tout l'intérêt du système SIBYLLE. *SibyMot* doit répondre à deux exigences contradictoires :

- son rôle est de présenter à la personne handicapée une liste de prédictions qui doit être ordonnée par vraisemblances décroissantes. Ce classement ne peut être obtenu que par des méthodes numériques. De ce point de vue, les modèles stochastiques de langage répondent à nos besoins,
- à l'opposé, les systèmes statistiques de prédiction ascendante présentent des performances assez médiocres (cf. § 2.3.2). Le caractère local et superficiel de l'analyse qu'ils mettent en œuvre leur interdit une prédiction plus précise. On retrouve sur ce domaine précis des limitations bien connues des modèles markoviens (N-grammaires). Le recours à des traitements linguistiques plus fouillés semble donc nécessaire à la mise en place d'une prédiction efficace.

Ainsi, la question spécifique de la prédiction lexicale pour l'aide à la communication relève d'un problème important qui intéresse l'ingénierie des langues dans sa globalité. Il s'agit de la mise en place de techniques de TAL statistique à fort ancrage linguistique. Certains chercheurs, tels Martin Rajman à l'EPFL, abordent cette question par la voie des grammaires stochastiques[86]. Je propose au contraire de rester dans un cadre markovien en basant les modèles de langage sur une information linguistique plus profonde que les simples cooccurrences de mots.

Mon idée est de combiner les approches stochastiques avec les niveaux de description qui sont généralement utilisés en TAL robuste (*shallow parsing*). Aussi n'est-il pas étonnant de retrouver dans *SibyMot* des principes rencontrés avec les systèmes de compréhension ROMUS et LOGUS :

- stratégie d'analyse incrémentale pour une prédiction robuste et efficace,
- importance du *chunk* comme unité d'analyse et de prédiction statistique.

A ma connaissance, cette proposition est totalement originale dans le domaine du TAL statistique. Seul le modèle structurel de Chelba et Jelinek[87], qui repose également sur une analyse en groupes noyaux, s'en rapproche. La figure 4.6 présente l'architecture générale de *SibyMot*.

---

[86] Rozenknop A., Chappelier J.-C., Rajman M. (2003) Apprentissage discriminant pour les grammaires à substitution d'arbres. actes TALN'2003, Batz-sur-Mer, France. 225-234

[87] Chelba C., Jelinek F. (2000) Structured language modeling. *Computer Speech and Language*, 14(4), 283-332.

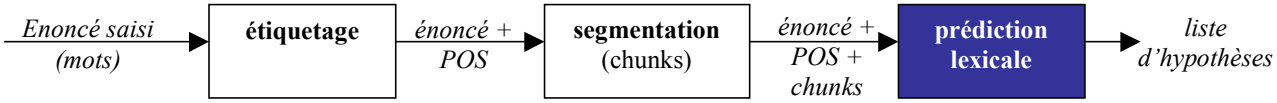

Figure 4.6 *— Architecture du système SibyMot*

La prédiction lexicale s'effectue en deux étapes. On procède tout d'abord à une analyse de l'énoncé en cours de saisie. Cette étape fournit la liste des parties du discours (POS) associées à chaque mot ainsi qu'une segmentation en *chunks* de l'énoncé. Ces indices linguistiques vont ensuite nourrir le modèle markovien de prédiction que nous avons élaboré.

Etiquetage — L'étiquetage de l'énoncé en parties du discours repose sur l'utilisation d'un modèle de langage trigramme. Ce modèle repose sur l'hypothèse réductrice que la probabilité d'affectation d'une étiquette C_i à un mot W_i dépend de la probabilité d'émission de ce mot (en fait, du lemme dans le cas de Sybille) pour la catégorie C_i et de la probabilité d'observer cette catégorie connaissant les deux catégories précédentes de l'énoncé. Soit pour l'ensemble de l'énoncé[88] :

$$Etiq_{solution} = \operatorname{argmax} \sum_{i} Pr(W_i \mid C_i) \,.\, Pr(C_i \mid C_{i-1}, C_{i-2})$$

Comme l'a montré la campagne d'évaluation GRACE (cf. chapitre 1, § 3.2), d'autres techniques peuvent être envisagées, qui présentent certainement de meilleurs performances. Notre choix répond en fait à des considérations très pragmatiques :

- les personnes handicapées possèdent généralement des ordinateurs dont la puissance est limitée. Une analyse à base de trigramme représente un bon compromis entre robustesse et combinatoire,
- comme je l'ai déjà noté au chapitre premier, les approches stochastiques constituent également un bon compromis entre coût de mise en œuvre et efficacité. Etant donné que la prédiction lexicale peut supporter les erreurs d'étiquetage d'une manière relativement transparente[89], nous avons fait le choix d'une conception rapide du module d'étiquetage. Rien ne nous empêchera de revenir ultérieurement à des solutions plus performantes.

L'étiquetage repose sur un jeu de catégories grammaticales qui s'inspire de celui défini pour l'action GRACE, à la différence importante que nous n'intégrons pas d'information morphosyntaxique pour réduire la taille du modèle. L'apprentissage du modèle markovien a été réalisé sur un extrait du journal *Le Monde* étiqueté préalablement par le logiciel Cordial Analyseur de la société Synapse.

L'étape suivante de segmentation ne travaille pas sur la meilleure hypothèse d'étiquetage mais sur un graphe d'hypothèses probabilisées. L'influence des erreurs d'étiquetage s'en trouve par conséquent réduit, ce qui limite également la criticité du choix de la technique d'étiquetage.

Segmentation — Nous avons choisi comme unité de découpage des chunks comparables à ceux qui sont manipulés par les systèmes LOGUS et ROMUS.

La structure des chunks observés a tout d'abord été décrite par un ensemble de patterns locaux (grammaire des chunks) que l'on peut identifier à des expressions régulières. Ces patterns reposent sur les catégories grammaticales et non pas sur les mots eux-mêmes. Par exemple :

GN ::= DET NC ADJ

La segmentation repose sur une analyse statistique. On affecte en effet à chaque pattern une probabilité d'émission suivant une procédure d'apprentissage par renforcement non supervisé.

---

[88] Pour une justification détaillée de cette formule : Paroubek P., Rajman M. (2000) Etiquetage morpho-syntaxique, In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection *IC2*, Hermès, Paris. 131-150.
Dans *SibyMot*, ce n'est pas le mot mais son lemme qui tient lieu de W_i. Ce choix est une réponse computationnelle aux contraintes que pose la puissance limitée des ordinateurs embarqués sur les fauteuils des handicapés.

[89] Une mauvaise prédiction se traduit au pire par des propositions erronées de la part du système, mais jamais par un comportement bloquant pour l'utilisateur.

L'analyse consiste alors à rechercher la segmentation de plus forte probabilité à l'aide d'un modèle markovien reposant sur deux paramètres :

1) la probabilité d'émission d'une succession de mots $W_{i,1}$ … $W_{i,j}$ (en fait, de lemmes dans le cadre de Sibylle) de parties du discours $C_{i,1...}$ $C_{i,j}$ pour chaque chunk G_i,

2) la probabilité d'émission du chunk G_i connaissant les deux chunks précédents G_{i-1} et G_{i-2}.

Formellement, la meilleure segmentation est donc donnée par la formule:

$$Seg_{solution} = \text{argmax} \sum_{i,j} Pr(C_{i,1...}\ C_{i,j} \mid G_i)\ .\ Pr(G_i \mid G_{i-1}, G_{i-2})$$

En pratique, cette étape fournit en sortie un graphe probabilisé d'hypothèses de segmentation.

Prédiction — *SibyMot* va utiliser l'ensemble des informations élaborées par les modules précédents (segmentation en chunk et parties du discours) pour fonder ses prédictions lexicales. Nous sommes donc en présence d'un analyseur statistique sensiblement plus profond que les modèles markoviens classiques, puisque la structure de l'énoncé est partiellement considérée par le modèle.

Relativement complexe, le processus de prédiction est réalisé en cinq étapes (figure 4.7) :

1) Prédiction des chunks à venir

2) Prédiction des relations de dépendances entre ces chunks,

3) Prédiction des lemmes à venir,

4) Prédiction des flexions des mots à venir, indépendamment des lemmes prédits,

5) Prédiction des mots à partir des informations sur les lemmes et les flexions.

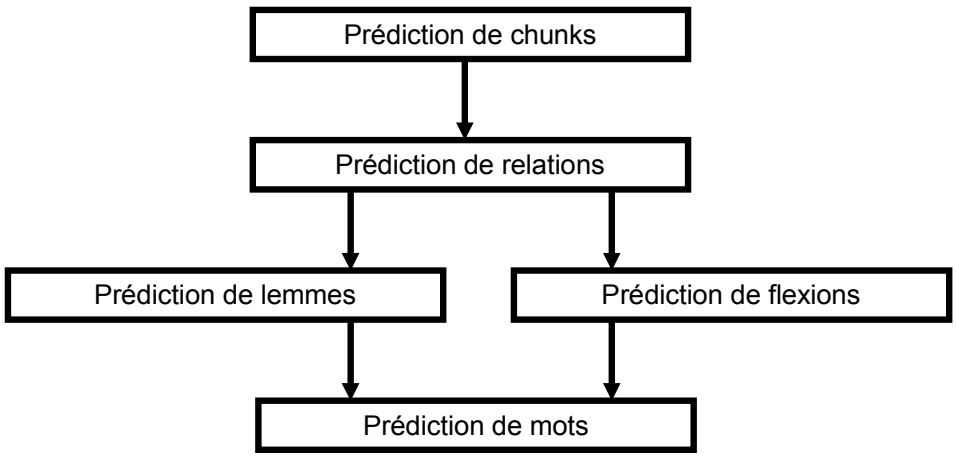

***Figure 4.7** — Architecture du module de prédiction du système SibyMot*

Je n'entrerai pas ici dans les détails de la réalisation du système[90]. Expliquons simplement les principes sous-jacents à ce modèle structurel de prédiction. Chaque étape fait appel à une modélisation markovienne de type tri-gram.

1) **Prédiction des chunks** — La première étape de la prédiction est chargée de fournir l'ensemble des segmentations possibles pour le mot à venir. Elle s'appuie sur la même grammaire des chunks que l'analyseur et donne une estimation de la probabilité de chacune des segmentations. Ces segmentations peuvent correspondre à deux situations : le mot à venir continue un chunk en cours ou marque au contraire le début d'un nouveau chunk dont on prédit également la catégorie

[90] Le lecteur intéressé pourra consulter : Schadle I. (2003) Sibylle : système linguistique d'aide à la communication pour les personnes handicapées. Doctorat Université de Bretagne Sud, Vannes, France. 18 décembre 2003. Rapport de recherche VALORIA-CORAIL-2003-03.

syntagmatique.

2) **Prédiction des relations entre chunk** — L'objectif de cette étape essentielle est de mettre en relation le(s) dernier(s) chunk(s) prédit(s) à l'étape précédente avec les *n-1* chunks précédents (N = 3 dans le cas du modèle markovien utilisé par Sybille). Le but implicite de cette étape est de capter, au moins partiellement, les relations de dépendance syntaxique entre syntagmes[91]. Pour réaliser cette mise en relation, on utilise uniquement l'étiquette syntagmatique attribuée aux chunks. Notons que le modèle ne cherche pas à typer les relations caractérisées.

À ce stade, la prédiction dispose de toutes les informations structurelles nécessaires pour estimer les probabilités d'apparition de chaque mot du lexique. Afin de réduire l'espace de recherche du modèle, la prédiction des mots est réalisée en deux étapes parallèles et indépendantes, qui portent respectivement sur les lemmes attendus et sur les catégories de flexion (genre, nombre, personne) qu'on doit leur appliquer pour obtenir les mots prédits.

3) **Prédiction des lemmes** — L'estimation des lemmes combine une probabilité tri-lemme (à l'image du n-gramme) et une probabilité fondée sur les têtes de chunks et leurs relations. Sur la segmentation [*l'année*]$_{GN}$ [*du singe*]$_{GP}$ [*a* < **??Participe passé**>]$_{GV}$ dont on a caractérisé la relation GV-GN, ces informations structurelles permettront d'obtenir des probabilités fortes pour les verbes associés au mot *année*. Le système proposera des lemmes tels que *commencer*, *débuter*, etc.

4) **Prédiction des flexions** — Parallèlement à l'estimation des lemmes, cette étape délivre une estimation de flexion pour chaque segmentation. Grâce aux relations de dépendances établies précédemment, une prise en compte des accords entre chunks est rendue possible.

5) **Prédiction des mots** — Les probabilités fournies par les deux étapes précédentes sont enfin combinées pour estimer la probabilité d'apparition de chaque mot du lexique de SibyMot.

A fin de prédiction, *SibyMot* affiche une liste d'hypothèses classées par probabilités décroissantes. La taille de la liste de prédiction est paramétrable (entre une et dix formes fléchies). Je vais tenter de montrer sur un exemple d'utilisation la puissance de cette prédiction.

2.4.4. SibyMot : un exemple d'utilisation

Nous allons considérer ici une session d'utilisation avec une liste de prédiction de 5 hypothèses. Supposons que l'on veuille précisément écrire la dernière phrase du paragraphe précédent :

(4.6) *je vais tenter de montrer sur un exemple d'utilisation la puissance de cette prédiction*.

En début de phrase, *SibyMot* propose une première liste d'hypothèses: *le, la*, *un*, *les*, *il*. Si tous ces termes sont bien susceptibles d'ouvrir l'énoncé, ceux-ci ne correspondent pas au mot recherché. L'utilisateur saisit donc la première lettre « j ». Les prédictions lexicales fournies par *SibyMot* sont filtrées pour ne conserver que les mots commençant par cette lettre : le pronom *je* apparaît en troisième position (figure 4.8a). L'utilisateur sélectionne cette proposition.

Cinq termes sont alors proposés par le système : *ne*, *n'*, *ai*, *suis*, *se*. On observe que les hypothèses verbales sont correctement accordées avec le pronom. Ces propositions ne correspondent pas aux attentes de l'utilisateur, qui saisit la première lettre du mot recherche : « v ». Après filtrage, *SibyMot*, le verbe attendu est en tête des prédictions (figure 4.8b) .

[91] En ce sens, notre système partage les mêmes objectifs que le modèle structurel de Chelba et Jelinek (2000) *op. cit*.

Figure 4.8 *— Exemple de prédictions effectuées par le module SibyMot. (a) à gauche, liste produite*

après la saisie de /j/ ; (b) au milieu, liste de prédiction après la saisie de /je v/ ; (c) à droite, liste obtenue après la saisie du message /je vais t/

Les cinq mots proposés après sélection du verbe ne répondent toujours pas aux attentes. Après saisie de la première lettre du mot suivant (« t »), on obtient cinq nouvelles prédictions : *très*, *toujours*, *tout*, *trop*, *trouver*. Tout en étant cohérentes, ces prédictions ne nous intéressent pas. On saisit donc la seconde lettre du mot *tenter*. Celui-ci est alors proposé en seconde position (figure 4.8c). Au passage, on remarque que les verbes proposés sont tous à l'infinitif. La construction « modal + verbe infinitif » a bien été modélisée par le système. Cette connaissance s'appuie essentiellement sur la probabilité du module de prédiction. Le mot suivant (*de*) est directement proposé en quatrième position de la liste de prédiction suivante. L'utilisateur le sélectionne donc directement.

La saisie se poursuit jusqu'à la fin de l'énoncé. Au final, 71% des saisies auront été économisées. Ce gain tient compte des saisies supplémentaires nécessitées par la sélection des items dans la liste de prédiction.

2.4.5. Validation expérimentale du système SIBYLLE

La dernière version de *SibyMot* a fait l'objet d'une évaluation sur un corpus de 50 000 mots extraits du corpus *Le Monde*. Le tableau 4.9 donne les résultats de cette évaluation, en terme de capacité d'économie de saisie, pour un fonctionnement avec une liste de 5 prédictions lexicales à chaque saisie. Le système a été comparé avec différents modèles N-grams (N= 1, 2 et 3) utilisés avec le même mode de fonctionnement.

Tableau 4.9 *— Performances comparées de SibyMot et de modèles markoviens classiques : gain de saisie (proportion de sélections évitées par rapport à une saisie sans assistance)*

Modèle	1-gram	2-gram	3-gram	SibyMot
Economie de saisie	43,9 %	51,2 %	55,8 %	57,1 %

On relèvera tout d'abord que le taux d'économie de saisie atteint par SibyMot (57,1%) est très encourageant et soutient la comparaison avec les systèmes de recherche déjà développés. Or, la comparaison avec les modèles markoviens montre que cette capacité de prédiction provient de l'intégration d'informations structurelles (syntaxiques). En effet, *SibyMot* présente des performances supérieures au modèle tri-gram, alors que chaque étape du système repose sur une modélisation stochastiques équivalente.

Afin de mieux étudier les capacités du système, j'ai réalisé une petite étude portant sur 12 énoncés différents (tableau 4.10). Suivant une démarche analogue aux évaluations DCR ou DEFI, cette expérience vise une analyse discriminante à son échelle. Elle a en effet porté sur différents modes d'utilisation (liste de 1, 5 ou 10 mots prédits) ainsi que sur quatre genres différents :

- texte journaliste (extraits du journal Libération),
- courrier (énoncés comportant en particulier des formules types d'ouverture ou de clôture),
- communication orale de nécessité ou de la vie quotidienne,
- travail scolaire (réponses écrites à un problème arithmétique ou travail sur fiche de lecture)

L'évaluation a porté une fois encore sur l'économie d'opérations de saisie, et non pas sur la vitesse de rédaction des énoncés. En effet, cette deuxième variable dépend fortement du handicap

considéré. Par ailleurs, c'est la répétition des saisies, plus que la lenteur de la communication, qui fatigue et décourage les personnes handicapées.

Tableau 4.10 *— SibyMot : gain de saisie sur différents énoncés (proportion de sélections évitées par rapport à une saisie sans assistance).*

Enoncé	1 mot	5 mot	10 mot
(1) Quatorze mois après leur prise de fonction, certains ministres sont déjà au bout du rouleau.	46 %	60 %	69 %
(2) Plusieurs ministres pensent qu'aucun mouvement n'aura lieu avant les élections	49 %	69 %	72 %
(3) A l'occasion des élections, certains ministres pourraient choisir de rester à la tête d'une région	43 %	63 %	67 %
(4) A la quasi-unanimité, l' assemblée nationale a voté l'instauration de la faillite civile pour lutter contre le surendettement qui frappe près de 600_000 familles	45 %	67 %	68 %
Moyenne texte journalistique	***45,8 %***	***64,8 %***	***69 %***
(5) A l'attention de M. le directeur des affaires sanitaires et sociales	46 %	68 %	70 %
(6) Je joins également à mon courrier les documents relatifs à la présente demande	46 %	59 %	64 %
(7) Veuillez agréer l'expression de mes sentiments les meilleurs	41 %	50 %	51 %
(8) J'aimerais bien changer d'émission	43 %	59 %	67 %
(9) Est-ce qu'on peut manger des légumes verts ce soir ?	42 %	55 %	62 %
(10) Peux-tu m'apporter le journal municipal ?	47 %	60 %	62 %
(11) Après la distribution, Jean a trois bonbons et cinq sucettes dans sa poche	42 %	52 %	56 %
(12) C'est Patrick qui a caché les chaussures neuves de Louise dans la grange	53 %	64 %	66 %
Moyenne autres genres	***45,0 %***	***58,4 %***	***62,3 %***
Moyenne	**45,3 %**	**60,5 %**	**65,3 %**

Le tableau 4.10 donne les résultats de l'expérience énoncé par énoncé. A chaque fois, on a relevé la proportion de saisies évitées grâce à la prédiction. Toute sélection dans la liste de prédiction lexicale est bien entendu comptabilisée comme une opération de saisie.

Plusieurs enseignements peuvent être tirés de cette expérience. Tout d'abord, on note que les performances de la prédiction lexicale sont très encourageantes. Même dans le cas difficile où la liste de prédiction se limite à un seul item, l'économie de saisie (45,3 %) égale les performances relevées dans la littérature[92]. Dans les autres cas (N= 5 ou N= 10), le gain de saisie est significatif. On remarque une tendance asymptotique qui nous conduit à penser qu'une liste de prédiction à cinq éléments constitue un bon compromis entre rapidité de composition des messages et effort de sélection. Cette observation doit cependant s'accorder à la situation de chaque handicapé.

L'analyse des résultats suivant le genre d'énoncé est également intéressante. Comme on pouvait s'y attendre, les meilleures performances sont enregistrées sur les textes journalistiques. L'influence du corpus d'apprentissage est donc significative. Cependant, la dégradation des résultats sur les autres genres reste relativement limitée. On remarque en particulier que :

- les performances globales du système restent acceptables en présence de mots hors vocabulaire. Le terme *sucettes* n'a pas été observé sur le corpus d'apprentissage extrait du journal *Le Monde*.

[92] Voir la remarque du paragraphe 2.3.2 : gains de saisie compris entre 27% et 46 % suivant les systèmes.

Quoique plus faible, le gain de saisie observé sur l'énoncé (11) reste pourtant correct,

- le système n'a pas été conçu pour apprendre des formules « type » telles que celles utilisées dans les courriers. Il reste cependant capable d'effectuer des prédictions locales efficaces. C'est ainsi que l'énoncé (7) bénéficie d'une économie de saisie encore appréciable, alors que ni la forme fléchie *veuillez* ni l'infinitif *agréer* n'a été observée au cours de l'apprentissage.

Ainsi, ces observations montrent que si on a intérêt à diversifier les corpus d'apprentissage, l'influence de ces données n'est pas réellement critique. *SibyMot* présente donc de solides atouts en matière de généricité.

2.4.6. Conclusion : valorisation et généralisation des travaux sur le handicap

Ces résultats montrent que le système SIBYLLE constitue une avancée significative dans le domaine de l'aide à la communication pour personnes handicapées. Les gains de saisie qu'autorise le module *SibyMot*, combinés à l'optimisation par *SibyLettre* du défilement du curseur sur le clavier simulé, permettent en théorie d'atteindre une vitesse de saisie quatre fois supérieure à celle des meilleurs systèmes actuellement commercialisés.

La dernière version du système SIBYLLE fait actuellement l'objet d'une validation écologique auprès de personnes handicapées du centre de Kerpape. Les premiers résultats de cette expérimentation sont très concluants : les patients s'adaptent sans effort au système et apprécient la rapidité de saisie permise par le système. Il n'est donc pas étonnant que SIBYLLE ait retenu l'attention de la société MicroVocal, qui travaille depuis de longues années dans le domaine de l'aide logicielle aux handicapés. Nous avons ainsi signé un accord de commercialisation du module de prédiction *SibyMot* avec cette société. Le module de prédiction sera ainsi intégré dès 2004 aux applications développées par la société. Il est à noter que MicroVocal réalise sous licence IBM les logiciels distribués par cette société dans le cadre de ces programmes Handicap. La distribution de *SibyMot* s'appuiera donc sur la puissance commerciale d'un groupe d'envergure mondiale. On peut donc espérer qu'un grand nombre de personnes handicapées pourront bénéficier, à un prix compétitif, des apports de ce système. Pour l'heure, cette commercialisation ne concerne que le français. Nous discutons actuellement du développement d'une version anglophone du système.

Cette valorisation ne doit pas masquer l'intérêt scientifique de nos travaux. La portée du système de prédiction lexicale dépasse en effet la seule problématique de l'aide au handicap. Le système *SibyMot* représente une tentative originale d'intégration de connaissances linguistiques relativement profondes dans un modèle de langage stochastique. A ce titre, il intéresse l'ingénierie des langues dans sa globalité et peut constituer une voie prometteuse pour la création de systèmes hybrides à l'interface entre le TAL probabiliste et le TAL robuste. On retrouve ici une des préoccupations qui guide mes recherches depuis de nombreuses années[93].

Les résultats prometteurs du module *SibyMot* nous questionnent ainsi sur l'extension de ce modèle de langage structurel à d'autres thématiques plus générales. En particulier, on peut s'interroger sur ses capacités de segmenteur de langue générale. Ce niveau de traitement intermédiaire entre *shallow parsing* et analyse syntaxique profonde sera prochainement étudié dans le cadre de la campagne d'évaluation EASy des analyseurs syntaxiques du français (action TECHNOLANGUE). Reposant sur le paradigme PEAS[94], cette évaluation étudiera la segmentation d'énoncés écrits ou oraux ainsi que la caractérisation des dépendances entre chunks. Une version adaptée de participera de *SibyMot* à cette campagne d'évaluation qui devrait nous fournir des indications sur la généricité du système.

---

[93] Cette question constituait par exemple la motivation principale de l'atelier « méthodes hybrides TALN / TALP » que j'ai organisé avec Damien Genthial dans le cadre du congrès TALN'1999.

[94] Gendner V., Illouz G., Jardino M., Monceaux L., Paroubek P., Robba I., Vilnat A. (2002) A protocol for evaluation analyzers of syntax (PEAS). Actes *3rd International Conference on Language Resources and Evaluation*, *LREC'2002*. Las Palmas de Gran Canaria, Espagne. 590-596.

3. CONCLUSION

La présentation de mes travaux sur la compréhension de la parole et l'aide au handicap montre qu'une même préoccupation scientifique (la mise en œuvre de traitements linguistiques robustes et détaillés) peut être conduite suivant des solutions relativement différentes (modèles probabilistes, automates pour une analyse superficielle de surface ou au contraire méthodes logiques avancées). Il n'en reste pas moins que tous mes travaux suivent une même approche. Tous les systèmes que j'ai présentés reposent en effet sur une stratégie d'analyse intégrant un premier étage de segmentation en chunks.

Comme le montrent nos travaux mais aussi d'autres recherches en TAL robuste, cette démarche constitue une solution séduisante pour combiner robustesse de traitement et finesse d'analyse. Elle nous permet également de capitaliser et de réutiliser des compétences sur des domaines d'applications relativement différents. Compte tenu de la taille modeste du groupe CORAIL, cette préoccupation ingénierique est essentielle à la réussite de nos travaux. Elle se retrouvera sans aucun doute dans mes recherches futures, pour lesquelles je vais maintenant esquisser quelques perspectives.

5. Conclusion :
TAL, linguistique et sciences cognitives

1. BILAN

Tout au long de ce document, j'ai cherché à montrer en quoi l'ingénierie des langues risquait de se perdre en oubliant ses fondamentaux, c'est-à-dire en ignorant le caractère spécifique du matériau sur lequel porte son étude. Cette inquiétude peut passer pour exagérée. A bien y regarder, on constate pourtant que les technologies langagières ont développé des pratiques et des méthodes que l'on retrouve à l'identique dans d'autres domaines relevant des sciences de l'ingénieur. Pour ne prendre qu'un exemple, quelle différence existe-t-il entre le modèle de Markov d'un système de reconnaissance de la parole et celui utilisé pour la modélisation de la houle en haute mer[1] ? Aucune en vérité. Le langage constitue pourtant une information très particulière, résultat de dizaines de milliers d'années d'évolution humaine. Son ancrage cognitif et social, sa complexité et son ambiguïté interdisent son identification à des observations météorologiques ou à des signaux de contrôle issus de n'importe quel procédé industriel. Bien souvent, ce sont pourtant les mêmes méthodes qui sont utilisées pour les analyser.

A l'heure où ces techniques se sont généralisées en traitement automatiques des langues, il me semble légitime de s'interroger sur les limites d'une recherche reposant sur des considérations exclusivement ingéniériques. A plusieurs reprises, j'ai tenté de montrer qu'une telle démarche pouvait être une source d'aveuglement nous conduisant à ignorer les besoins réels de l'utilisateur. Pour reprendre un lieu commun, nous risquons de développer des « solutions d'ingénieur », avec les connotations négatives que l'on sait attachées à ce terme.

Mon propos n'est pas de nier l'apport méthodologique et technologique qu'a constitué l'émergence de ces approches. Celles-ci ont révolutionné un domaine de recherche qui s'interrogeait sur ses échecs, au point qu'un retour en arrière ne saurait être désormais envisageable. Simplement, je pense que l'ingénierie des langues doit viser un plus fort ancrage linguistique si elle veut atteindre ses objectifs.

On pourrait objecter que l'ingénierie des langues a déjà pris conscience de ses limites et que je ne fais qu'enfoncer des portes ouvertes. Il est vrai que l'on peut observer, ça et là, une réflexion sur les méthodes de traitement actuellement utilisées. Un exemple caractéristique de cette prise de recul concerne la recherche de modèles de langages stochastiques qui intègrent des connaissances syntaxiques partielles[2]. Cette réflexion ne me semble cependant pas totalement à la mesure des enjeux qu'implique la généralisation des technologies langagières auprès du grand public.

Il me semble que cette carence concerne, avant tout, nos pratiques scientifiques. Je regrette en particulier que l'ingénierie des langues considère les corpus comme des données d'apprentissage (d'où le terme de *ressources* linguistiques) plus que comme un témoignage des usages langagiers. De même, si de nombreux chercheurs ont pointé le manque de pouvoir prédictif des évaluations globales de performances, on constate que la prédominance de ces paradigmes reste sans partage.

Ce sont donc nos pratiques quotidiennes qui doivent faire l'objet d'un aggiornamento. Dans le chapitre premier de ce mémoire, j'ai discuté des limites des méthodes de conception par bootstrap. J'ai proposé de revaloriser le rôle des analyses amonts de corpus pilotes, dont on ne peut que constater l'abandon au cours des dix dernières années. Le caractère prédictif des analyses d'usages que j'ai présentées démontrent pourtant l'intérêt de ce type d'études.

---

1 Monbet V., Marteau P.-F. (2001) Continuous Space Discrete Time Markov Models for Multivariate Sea State Parameter Processes. Actes *11th International Offshore and Polar Engineering Conference & Exhibition*. Stavanger, Norvège.

2 Chelba C., Jelinek F. (2000) Structured language modeling. *Computer Speech and Language*, 14(4), 283-332.

Au cours du second chapitre, j'ai proposé une alternative (méthodologies DCR et DEFI) aux limites des évaluations globales de performances. Les expériences réalisées avec ces nouveaux paradigmes discriminants montrent qu'un diagnostic fin du comportement des systèmes peut aider l'ingénierie des langues à détecter les principaux problèmes sur lesquels elle doit s'interroger. Comme par hasard, ce type d'évaluation ne peut s'envisager sans une réflexion amont sur les usages langagiers qui doivent être considérés. C'est pourquoi je me réjouis que les recherches sur les paradigmes DCR et DEFI rencontrent un écho croissant au sein de la communauté scientifique.

Cet ancrage linguistique se retrouve dans les réalisations du groupe de recherche CORAIL que je dirige. Qu'ils concernent la compréhension de la parole ou l'aide à la communication pour handicapés, ces travaux visent une analyse linguistique détaillée et robuste. Les techniques employées, qui relèvent de motivations linguistiques clairement assumées, n'ont rien d'ésotériques[3]. Au contraire, elles découlent de l'adaptation de méthodes déjà éprouvées dans d'autres contextes. Les performances encourageantes de nos systèmes sont autant d'appuis à cette démarche.

2. PERSPECTIVES DE RECHERCHE

J'ai bien conscience que ce plaidoyer en faveur d'une ingénierie des langues plus linguistique relève plus de l'analyse critique que de la démonstration. Les travaux du groupe CORAIL sont souvent trop récents pour imposer leurs résultats avec force. Notre équipe, issue de la plus jeune université de France, n'a que quelques années d'existence et a consacré une partie de ses efforts à se structurer de manière cohérente. J'espère que ce mémoire témoigne du chemin déjà parcouru, et qu'il montre que notre groupe aura à l'avenir les moyens de faire valoir l'originalité et la pertinence de son programme scientifique.

Aussi, mon objectif est que le groupe CORAIL soit en mesure de présenter des réalisations qui témoignent de la qualité des recherches que nous entreprenons. Compte tenu de l'évolution du traitement automatique des langues, cette démonstration ne peut s'inscrire que dans un cadre résolument ingénierique. Notre participation à plusieurs projets TECHNOLANGUE (campagnes d'évaluation MEDIA et EASy, projet OURAL) constitue ainsi une opportunité particulièrement motivante de démontrer notre savoir-faire.

Cette recherche de reconnaissance doit toutefois être atteinte sans que nous ayons à renier notre démarche scientifique. C'est pourquoi les perspectives que je vais esquisser s'intègrent dans la structuration actuelle des recherches menées par le groupe CORAIL. Cette organisation n'est en effet par le fruit des circonstances, mais bien le produit d'une réflexion que j'ai tenté d'exposer dans ce mémoire. C'est pourquoi je pense que le développement futur du groupe CORAIL se traduira avant tout par un approfondissement des quatre axes de recherche que j'avais définis :

- **Ressources linguistiques orales et linguistique de corpus** — J'ai souligné au cours du second chapitre l'importance que je comptais accorder à la constitution des corpus de dialogue oral dans les activités futures du groupe CORAIL. Ces travaux s'inscriront dans le cadre du projet PAROLE PUBLIQUE déjà en place. En cas de réussite, cette activité pourrait donner lieu à la création d'une ERT (Equipe de Recherche Technologique). Cette ERT jouerait un rôle de centre de ressources pour le dialogue oral destiné à l'ensemble de la communauté. C'est en tout cas l'objectif à cinq ans que j'ai fixé au projet de Plan Pluri-Formation (PPF « CORAIL ») que j'ai monté en collaboration avec le laboratoire CRELLIC/ADICORE. Cette proposition est en cours d'évaluation dans le cadre du renouvellement du contrat quadriennal de l'Université de Bretagne Sud. Par ailleurs, le projet PAROLE PUBLIQUE sera certainement amené à intégrer une dimension régionale portant sur la langue bretonne. Je travaille actuellement à la mise en place d'un projet « corpus breton » qui réunirait d'autres collègues des universités de Rennes 1 et 2.

  Je compte bien entendu poursuivre mes recherches sur les études amont de corpus pilotes. Cette

[3] Un bémol à cette affirmation pourrait être apporté dans le cas du système LOGUS. Si l'utilisation d'approches logiques en compréhension de parole est très originale, on remarquera tout de même que les travaux portant sur les grammaires catégorielles sont beaucoup moins rares en traitement du langage écrit.

activité sera malheureusement toujours limitée par le caractère monodisciplinaire du laboratoire VALORIA. C'est précisément pour répondre à l'absence de chercheurs en linguistique sur le site de Vannes que je travaille depuis plus de deux ans à un rapprochement du VALORIA avec le laboratoire CRELLIC/ADICORE (Université de Bretagne Sud, Lorient). Les centres d'intérêts principaux des deux laboratoires sont cependant différents. Mes recherches en linguistique de corpus risquent donc d'être guidées essentiellement par les besoins de conception que nous rencontrerons. Comme l'ont par exemple montré mes études sur les dislocations, cette contrainte n'empêche pas la production de résultats plus généraux. A l'avenir, je compte en tous cas étendre ces analyses à des études différentielles DHH / DHM afin d'étudier les variabilités entre des deux types de ressources complémentaires.

- **Evaluation des technologies langagières** — Les paradigmes de test DCR et DEFI ont eu une influence significative sur les recherches francophones en compréhension de la parole. La méthodologie PEACE, qui sera utilisée par la prochaine campagne MEDIA, répond ainsi à mes attentes en matière d'évaluation. La démonstration de l'intérêt d'une évaluation diagnostic n'est donc plus à faire. C'est pourquoi je compte essentiellement renforcer et approfondir ces travaux dans les années à venir. Tout d'abord, le groupe de travail « Compréhension » (GDR I3) que je dirige va poursuivre sur des phénomènes précis l'évaluation par défi des systèmes de compréhension. Il me semble par ailleurs essentiel de poursuivre une réflexion[4] sur l'évaluation diagnostic des systèmes de dialogue. Cette activité sera développée en parallèle à nos travaux sur le dialogue oral homme-machine (cf. infra).

  A l'opposé, un des mes objectifs prioritaires pour l'avenir concerne la mise en place de méthodes d'évaluation des systèmes d'aide à la communication. Les recherches sur le handicap souffrent en effet de l'absence de programmes d'évaluation qui permettraient de juger la pertinence des (trop ?) nombreuses approches qui fleurissent dans ce domaine. L'évaluation des technologies langagières d'aide au handicap nécessite une démarche résolument pluridisciplinaire : à côté des mesures brutes de performances, il est en effet essentiel d'intégrer des critères d'ergonomie, d'utilisabilité, d'adaptation au handicap qui ne relèvent pas des seules sciences de l'ingénieur. C'est pour répondre à ces interrogations pluridisciplinaires que j'ai proposé à Nadine Vigouroux de lancer une équipe projet dans le cadre du RTP « Handicap » du CNRS. Ce projet, d'une durée de trois ans, concernerait spécifiquement les systèmes d'aide à la communication langagière générale (cf. chapitre 4,§ 2.2). Dans mon esprit, il doit bien entendu donner lieu à évaluation diagnostic.

- **Dialogue oral homme-machine** — Le ton optimiste avec lequel je dresse le bilan de plus de sept années de recherche ne saurait cacher un regret : celui de n'avoir pu aborder de front la question du dialogue homme-machine en réalisant un système interactif complet. Cette situation s'explique par les efforts qu'ont nécessités la création ex nihilo du groupe CORAIL à mon arrivée à l'Université de Bretagne Sud, mais aussi par mon désir d'en assurer le développement progressif en toute autonomie scientifique. Avec les systèmes ROMUS et LOGUS, nous disposons désormais de briques de bases sur lesquelles asseoir la réalisation de systèmes interactif. Le grand chantier des années à venir devrait donc concerner le dialogue proprement dit. Il a d'ailleurs été entamé cette année dans le cadre des DEA de Julien Foulon (travail sur les co-références anaphoriques) et Frédéric Lamie (réalisation d'un serveur vocal R&D pour le renseignement aérien). Le développement d'un systèmes interactifs constitue cependant un effort conséquent qui ne saurait être mené en solitaire par une petite équipe. Je remarque à ce sujet que le nombre de dialogueurs opérationnels dans les centres de recherche français reste limité. C'est pourquoi ces avancées se réaliseront progressivement — en commençant par le calcul de la référence — en collaboration avec d'autres laboratoires. Cette année, une première proposition de co-encadrement de thèse avec l'IRISA/CORDIAL (Jacques Siroux) a ainsi été montée dans cet esprit. Je compte développer ces travaux sur le dialogue dans deux directions.

  Il m'apparaît tout d'abord intéressant de poursuivre cette recherche en conservant la démarche —

[4] Cette réflexion a déjà été ébauchée par le CLIPS-IMAG avec la méthodologie DCR (doctorat de M. Ahafhaf).

analyse superficielle incrémentale — qui préside à mes travaux actuels en compréhension de parole. Les premiers enseignements du stage de DEA de Julien Foulon, qui porte sur l'adaptation au langage parlé des techniques superficielles[5] de résolution des anaphores, me laissent cependant penser qu'il reste à mener une réflexion approfondie sur la généralisation du TAL robuste à ces niveaux de traitement.

A la suite de la thèse de Jeanne Villaneau, il me semble également intéressant de se pencher sur une question qui est largement ignorée par les concepteurs de systèmes d'information interactifs. Il s'agit de l'influence des techniques d'interrogation des bases de données sur la naturalité de l'interaction. Les requêtes SQL sont assez éloignées du comportement humain en matière de recherche d'information. En particulier, elles ne permettent aucune gradualité de la recherche (critères de préférences relatives). L'intégration de requêtes flexibles à base de prédicats flous avec une interprétation logique des marqueurs de préférence pourrait au contraire conduire à une interrogation plus naturelle. Cette approche serait compatible avec les traitements logiques mis en œuvre par le système LOGUS. Il me parait donc intéressant de travailler sur des modules de dialogue dont l'interface avec l'application autoriserait une interrogation plus naturelle. Ces travaux seraient réalisés en collaboration avec les équipes de l'IRISA qui s'intéressent à ces questions. Je pense bien entendu au projet LIS (*Logic Information Systems*) développé par Olivier Ridoux mais également aux travaux du projet BADINS (Patrick Bosc, LLI Lannion).

- **Dialogue médié par l'ordinateur : aide à la communication pour handicapés** — La réussite que connaît le système SIBYLLE nous incite à nous concentrer à court terme sur son développement : de nombreuses pistes d'amélioration, intéressantes d'un point de vue théorique, viennent à l'esprit, tandis que la commercialisation de ce logiciel posera certainement la question de son portage vers d'autres idiomes, dont l'anglais en premier lieu.

  La campagne d'évaluation EASy nous permettra prochainement de mieux appréhender le caractère générique de ces travaux. C'est, je pense, à la lumière de ces résultats que l'on pourra s'interroger sur l'extension à d'autres problématiques de notre modèle de langage structurel. L'application qui vient en premier à l'esprit concerne la prédiction linguistique pour organiseurs personnels et autres ordinateurs de poches, domaine où les enjeux économiques sont importants.

3. CONCLUSION : TAL ET SCIENCES COGNITIVES

Quelle que soit l'évolution de mes recherches futures, nul doute qu'elles seront toujours menées dans l'optique d'une ingénierie des langues à fort ancrage linguistique. Je compte donc poursuivre ce programme de recherche dans une perspective qui sera toujours résolument pluridisciplinaire.

L'interdisciplinarité est beaucoup moins encouragée en France qu'Outre-Atlantique. Elle m'apparaît pourtant comme une exigence salutaire en faveur d'une ouverture d'esprit que j'ai cherché à transmettre dans ce mémoire. Je tiens ainsi à m'en tenir à une attitude ouverte sur l'ensemble des champs d'investigations des sciences cognitives, en particulier dans le cadre de mon engagement à la direction de la revue *In Cognito – Cahiers Romans de Sciences Cognitives*.

Par delà la question de la confrontation des idées entre Intelligence Artificielle et Sciences du langage, les recherches en sciences cognitives ont souvent été une source d'inspiration féconde pour mes travaux. Au moment où l'ingénierie des langues se rapproche d'applications destinées au grand public, il me semble que la nécessité de ce regard sur la cognition humaine et sociale ne devrait qu'être plus manifeste.

---

[5] Mitkov R. (1998) Robust pronoun resolution with limited knowledge. Actes *36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, *ACL-COLING'98*, Montréal, Canada, 869-875.

Bibliographie

Abeillé A.(1993) Les nouvelles syntaxes : grammaires d'unification et analyse du français, Armand Colin, Paris.

Abeillé A., Clément L., Reyes R. (1998). TALANA annotated corpus : the first results. *1st Conference on Linguistic Resources and Evaluation*, *LREC'1998*, Grenade, Espagne, 992-999.

Abeillé A., Blache P. (2000) Grammaires et analyseurs syntaxiques. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 51-76.

Abeillé A., Clément L., Kinyon A. (2000) Bulding a treebank for French. Actes *2nd Conference on Linguistic Resources and Evaluation*, *LREC'2000*, Athènes, Grèce, 87-94.

Abney S. (1991) Parsing by chunks. *In* Berwick R., Abney S. and Tenny C. (Eds.) *Principle-based parsing*. Kluwer Academic Publ., Dordrecht, Pays-Bas. Disponible sur la Toile : www.sfs.nphil.uni-tuebingen.de/~abney

Abney S. (1996) Partial parsing via finite-state cascades. Actes *ESSLI'1996 Robust Parsing Workshop*. Disponible sur la Toile : www.sfs.nphil.uni-tuebingen.de/~abney.

Abraham M. (2000) Reconstruction de phrases oralisées à partir d'une écriture pictographique. Actes *Handicap'2000*, Paris, France. 151-156.

Adda G., Mariani J., Paroubek P., Rajman M. et Lecomte J. (1999) L'action GRACE d'évaluation de l'assignation des parties du discours pour le français, *Langues*, 2(2), 119-129.

Ait-Mokhtar S., Chanod J.-P., Roux C. (2002) Robustness beyond shallowness : incremental deep parsing. *Natural Language Engineering*, 8(2-3). 121-144.

Aït-Mokhtar S., Hagège C., Sándor Á. (2003) Problèmes d'intersubjectivité dans l'évaluation des analyseurs syntaxiques. Actes *Atelier TALN'2003 sur l'évaluation des analyseurs syntaxiques*. Batz-sur-Mer, France, Vol. 2, 57-66.

Alexandersson J., Reithinger N., Maier E. (1997) Insights into the dialogue processing of VERBMOBIL, Actes *5th Conference on Applied NLP*, *ANLP'97*, Washington, DC, 33-40.

Allen J., Byron D., Dzikovska M. (2000) An architecture for a generic dialogue shell. *Natural Language Engineering*. 6 (3-4). 213-228.

Allen J.F., Miller B.W., Ringger E.K., Sikorski T. (1996) Robust understanding in a dialogue system. Actes *34th Annual meeting of the Association for Computational Linguistics, ACL'96*. San Francisco, CA, 62-70.

ALPAC : Automatic Language Processing Advisory National Research Council (1966). Language and machines ; computers in translation and linguistics, rapport 1416, National Academy of Sciences, Washington, Etats-Unis.

Antoine J.Y. (1994) Coopération syntaxe-sémantique pour la compréhension de la parole spontanée. Thèse de Doctorat. INP Grenoble, Grenoble, France.

Antoine J.Y. (1995) Conception de dessins et Communication Homme - machine : améliorer l'interaction orale au niveau linguistique, In Zreik K., Caelen J. (ed.), *Communication en conception*, EuropIA éditions, Paris, France.

Antoine J.-Y., Bousquet-Vernhettes C., Goulian J., Kurdi M. Z., Rosset S., Vigouroux N., Villaneau J. (2002) Predictive and objective evaluation of speech understanding: the "challenge" evaluation campaign of the I3 speech workgroup of the French CNRS. Actes *3rd Conference on Language Resources and Evaluation*, *LREC'2002*, Las Palmas de Gran Canaria, Espagne.

Antoine J-Y., Caelen J. (1999) Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR (Demande, Contrôle, Résultat), *Langues*, 2(2). 130-139.

Antoine J.-Y., Genthial D. (1999), *Méthodes hybrides issues du TALN et du TAL Parlé : états des lieux et perspectives*, TALN'1999, atelier thématique « Méthodes hybrides pour le TAL robuste », Cargèse, France, 1-17.

Antoine J-Y., Goulian J. (2001) Etude des phénomènes d'extraction en français parlé sur deux corpus de dialogue oral finalisé. Application à la communication orale homme - machine. *Traitement Automatique des Langues*, *TAL*, 42(2), 413-440.

Antoine J.Y., Le Pévédic B. (2001) Ingénierie des langues et handicap. Actes *TALN'2001*, conférence associée *Ingénierie des Langues et Handicap*, Tours, France. vol. 2, 179-182.

Antoine J.-Y., Letellier-Zarshenas S., Nicolas P. , Schadle I., Caelen J. (2002) Corpus OTG et ECOLE_MASSY : vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement. Actes *TALN'200*2, Nancy, France. vol. 1, 319-324.

Antoine J-Y., Siroux J., Caelen J., Villaneau J., Goulian J., Ahafhaf M. (2000) Obtaining predictive results with an objective evaluation of spoken dialogue systems : experiments with the DCR assessment paradigm, Actes *2nd Conference on Language Resources and Evaluation*, *LREC'2000*, Athènes, Grèce.

Antoine J-Y. *et al*. (2001) Synthèse de la réunion d'analyse des résultats de la campagne d'évaluation par défi. Disponible sur la Toile: www.univ-ubs.fr/valoria/antoine/gdri3/Oct01.html

ARPA (1995). Proceedings of *1995 ARPA spoken language systems technology workshop*, Austin, Texas.

Aust H., Oerder M., Seide F., Steinbiss V. (1995) The Phillips automatic train timetable information system. *Speech Communication*, 17. 249-262.

Bahl L., Mercer R. (1976) Part of speech assignment by a statistical decision algorithm. Actes *International symposium on Information Theory*.

Baker J.K. (1975) Stochastic modeling for automatic speech understanding. In Reddy D.R. (Ed.) *Speech recognition*, Academic Press, New-York, NJ. 521-542.

Bangalore S. (1999) Suppertagging : an approach to almost parsing. *Computational Linguistics*, 25(2). 237-265.

Bar-Hillel Y. (1964). Language and information. Addison-Wesley, Reading, Etats-Unis

Barras C. *et al*. (1998) Transcriber : a free tool for segmenting, labeling and transcribing speech, Actes *1st Conference on Language Resources and Evaluation*, *LREC'98*. Grenade, Espagne,, pp. 1373-1376.

Basili R., Zanzotto F.M (2003) Parsing engineering and empirical robustness. *Natural Language Engineering*, 8 (2-3) 147-169.

Bates M., Boisen S., Makhoul J. (1992) Developing an evaluation methodology for spoken language systems. Actes *DARPA Specch and Natural Language Workshop*. 102-108.

Baum L.E., Eagon J.A. (1967) An inequality with applications to statistical estimation for probalistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*. 73, 360-363.

Bear J., Dowding J., Shriberg E. (1992) Integrating multiple knowldege sources for detection and correction of repairs in human-computer dialog. Actes *30th Annual Meeting of the ACL*, *ACL'92*. Newark, Etats-Unis. 56-63.

Bear J., Dowding J., Shriberg E., Price P. (1993) A system for labeling self-repairs in speech. *SRI Technical note S22.*

Béchet F., Nasr A., Spriet T., De Mori R. (1999) Modèles de langage à portée variable : application au traitement des homophones. Actes *TALN'1999*, Cargèse, France, 35-44.

Belleannée C., Brisset P., Ridoux O. (1999) A pragmatic reconstruction of λProlog. *Journal of Logic Programming*, 41(1). 67-102.

Bennacef S. (1995) Modélisation du dialogue oral homme - machine : mise en œuvre dans une application de demande d'informations. Thèse de Doctorat, Université Paris XI, Orsay, France.

Bennacef S., Devillers L., Rosset S., Lamel L. (1996) Dialog in the RailTel telephone-based

system. Actes *4th International Conference on Spoken Language Processing, ICSLP'1996*, Philadelphie, Pa. 550-553.

Benzitoun C., Caddeo S. (2002) La recherche automatique des appositions, Actes *2èmes journées de la linguistique de corpus*, Lorient, France, p. 8 (résumé).

Bernard P., Bernet C., Dendien J., Pierrel J.-M., Souvay G., Tucsnak Z. (2001) Ressources linguistiques informatisées de l'ATILF. Actes TALN'2001, Tours, France. vol 1, 333-338.

Bernard P., Dendien J., Lecomte J., Pierrel J.-M. (2002) Un ensemble de ressources informatisées et intégrées pour l'étude du français : FRANTEXT, TLFi, Dictionnaires de l'Académie et logiciel Stella. Actes *TALN'2002*, Nancy, France. 3-36.

Berrendonner A., Reichler-Béguelin M.J. (1995) Accords associatifs, *Cahiers de praxématique*, 24, 14-21Bessac M., Caelen G. (1995), Analyses pragmatiques, prosodiques et lexicales d'un corpus de dialogue oral homme - machine, Actes *JADT'95*, Rome, Italie, 363:370.

Biber D. (1988) Variations across speech and writing. Cambridge University Press, Press, Cambridge, MA.

Biber D. (1986) Spoken and written textual dimensions in English : resolving the contradictory findings. *Language*, 62(2), 384-414.

Bilange E. (1992) Dialogue personne-machine : modélisation et réalisation informatique. Hermès, Paris, France.

Bilger M. (2000) Petite typologie des conventions de transcription de l'oral : quelques aspects pratiques et théoriques. *Cahiers de l'Université de Perpignan*, n° 31, Presses Universitaires de Perpignan, Perpignan, France. 77-92.

Blanche-Benveniste C. (1997) Approches de la langue parlée en français, Coll. *L'essentiel Français*, Ophrys, Paris, France.

Blanche-Benveniste C., Bilger M., Rouget C., Van den Eynde K. (1990) Le français parlé : études grammaticales. CNRS, Paris, France.

Blanche-Benveniste C., Jeanjean C. (1987) Le français parlé : transcription et édition. Paris, Didier Erudition.

Blanche-Benveniste C., Rouget C., Sabio F. (2002) Choix de textes de français parlé : 36 extraits. Honoré Champion, Paris.

Boehm B. W. (1988) A spirale model of software development and enhancement. *IEEE Computer*, 21(5). 61-72.

Boehm B.W., Gray T.E., Seewaldt T. (1984) Prototyping versus specifying : a multi-project experiment. *IEEE Transactions on software engineering*, 10(3). 290-303.

Boissière P., Dours D. (2001) Comment VITIPI, un système d'assistance à l'écriture pour les personnes handicapées peut offrir des propriétés intéressantes pour le TALN. Actes *TALN'2001*, conférence associée *Ingénierie des langues et handicap*, Tours, France. vol. 2, 183-192.

Boissière P., Dours D. (2002) Vers un modèle d'aide à l'évaluation de systèmes d'assistance à l'écriture : application à VITIPI. Actes *Handicap'2002*.

Boite R., Bourlard H., Dutoit H., Hancq J. et Leich H. (2000), Traitement de la parole, Coll. Electricité, Presses Polytechniques et Universitaires Romandes, Lausanne, Suisse.

Bonhomme P. (2000) Codage et normalisation de ressources textuelles. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 173-192.

Bonneau-Maynard H., Devillers L. (2000) A framework for evaluating contextual understanding. Actes *6th International Conference on Spoken Language Processing, ICSLP'2000*. Pékin, Chine.

Bonneau-Maynard H., Devillers L., Rosset S. (2000) Predictive performance of dialog system. Actes *2nd International Conference on Language Resources and Evaluation*, *LREC'2000*, Athènes, Grèce.

Bonneau-Maynard H., Gauvain J.-L., Goodine D., Lamel L., Polifroni J., Seneff S. (1993) A French version of the MIT-ATIS system : portability issues. Actes *3rd European Conference on Speech Communication*, *Eurospeech'93*, Berlin, Allemagne.

Bouillon P. (1998) Traitement automatique des langues naturelles. Duculot, Bruxelles, Belgique.128-131.

Bouillon P., Fabre C., Sébillot P., Jacqmin L. (2000) Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL*, vol. 41 n° 2. 447-472. Hermès, Paris, France. 367-393.

Bourlard H., Morgan N. (1994) Connectionist speech recognition : a hybrid approach. Kluwer Academic Publ., Dordrecht, Pays-Bas.

Bousquet-Vernhettes C. (2002) Compréhension robuste de la parole spontanée dans le dialogue oral homme - machine : décodage conceptuel stochastique. Thèse de l'Université Paul Sabatier, Toulouse, France, 26 septembre 2002.

Brangier E., Gronier G. (2000) Conception d'un langage iconique pour grands handicapés moteurs aphasiques. Actes *Handicap'2000*, IFRATH. Paris, France. 93-100.

Bridle J.S. (1990) Alpha-nets : a recurrent " neural " network architecture with a Hiden Markov Model interpretation. *Speech Communication*, 9(1), 83-92.

Brill E. (1992) A simple rule-based part of speech tagger. *Actes 3rd Conference on Applied Natural Language Processing*, *ANLP'1992*, Trente, Italie.

Brill E. (1995) Transformation-based error-driven learning and natural language processing : a case study in part of speech tagging, *Computational Linguistics*, 21(4), 543-565.

Brown P., Cocke J., Pietra S. D., Jelinek F., Lafferty J.D., Mercer R.L. et Rossin P.S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2), 79-85.

Bruce B. (1975) Case systems for natural language. *Artificial Intelligence*. 6. 327-360.

Caelen J., Garcin P., Wret J., Reynier E. (1991) Interaction multimodale autour de l'application ICP-Draw. Actes *IHM'91*. Douradan, France. 1-12.

Caelen J. (1994) *Modélisation de la tâche et de l'usager*. Rapport de recherche projet DALI, GDR-PRC Communication Homme - machine.

Caelen J., Nasri M.K., Reynier E., Tattegrain H. (1990) Architecture et fonctionnement du système DIRA. De l'acoustique aux niveaux linguistiques. *Traitement du signal*, 7 (4). 345-366.

Caelen J., Zeiliger J., Bessac M., Siroux J., Perennou G. (1997) Les corpus pour l'évaluation du dialogue homme - machine. Actes des *1ères Journées Scientifiques et Techniques FRANCIL*, *JST'1997*, Avignon, France, 215-222. Texte repris dans : Chibout K., Mariani J., Masson N., Néel F. (Dir.) (2000) Ressources et évaluations en ingénierie des langues. De Boeck Université, Duculot, Bruxelle, Belgique. 417-435.

Carré R., Descout R., Eskénazi M., Mariani J., Rossi M. (1984). The French language database : defining, planning and recording a large database. Actes *1984 International Conference on Acoustics, Speech and Signal Processing*, *ICASSP'1984*. San Diego, CA, Etats-Unis. Vol. 3. 42-10.1 - 42.10-4.

Carroll J., Briscoe T. (1996) Robust parsing : a brief overview. Actes *ESSLI'1996 Robust Parsing Workshop*. Disponible sur la Toile : www.cogs.susx.ac.uk/lab/nlp/carroll/papers/esslli96.pdf.

Chanod J.P. (1994) Developpements en analyse syntaxique automatique. Actes TALN'1994, Marseille. France. 87-91.

Chappelier J.C., Rajman M., Aragües R., Rozenknop A. (1999). Lattice parsing for speech recognition. Actes *TALN'1999*, Cargèse, France. 95-104.

Charolles M. (2002) La référence et les expressions référentielles en français. Ophrys, Gap, France.

Chelba C., Jelinek F. (2000) Structured language modeling. *Computer Speech and Language*, 14(4), 283-332.

Chomsky N. (1957) Syntactc Structures, Mouton & Co., La Haye, Pays-Bas.

Chung G., Seneff S., Hetherington L. (1999) Towards Multi-Domain Speech Understanding Using a Two-Stage Recognizer, Actes *6th European Conference on Speech Communication and Technology*, *Eurospeech'1999*, Budapest, Hongrie. 2655-2658.

Church K. (1988). A stochastic parts program and noun phrase parser for unrestricted text, actes *Conference on Applied Natural Language Processing*, *ACL'1988*, Austin, TX, 136-143.

Church W.K., Mercer R. L. (1993) Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1), 1-25.

Clementino D., Fissore L. (1993) A man-machine dialogue system for speech access to train timetable information. Actes *3rd European Conference on Speech Communication and Technology*. *Eurospeech'93*. Berlin, Allemagne. 1863-1866.

Cohen J., Gish H. et Flanagan J. (1994) Switchboard, the second year. Actes *CAIP summer workshop in speech recognition: frontiers in Speech Processing II*.

Colineau N., Caelen J. (1997) Analyses de dialogues oraux et modélisation des actions de communication. Actes des *1ères Journées Scientifiques et Techniques FRANCIL*, *JST'1997*, Avignon, France, 447-454 ; Texte repris dans : Chibout K., Mariani J., Masson N., Néel F. (Dir.) (2000) Ressources et évaluations en ingénierie des langues. De Boeck Université, Duculot, Bruxelle, Belgique. 463-482.

Collectif (1993) Special issue on using large corpora :I. *Computational Linguistics*, 19(1), mars 1993.

Core M. G., Schubert L. K (1999) A syntactic framework for speech repairs and other disruptions, Actes *37th Annual Meeting of the Association for Computational Linguistics, ACL'1999.*

Coret A., Kremer P., Landi B., Schibler D., Schmitt L., Viscogliosi N. (1997) Accès à l'information textuelle en français : le cycle exploratoire Amaryllis. Actes *1ères Journées Scientifiques et Techniques du réseau FRANCIL*, *JST-FRANCIL'97*, Avignon, France. 5-8.

Covington M.A. (1990) A dependancy parser for variable-order languages, rapport de recherche AI-1990-01, University of Georgia, Etats-Unis.

Covington M. (1990) Parsing discontinuous constituents in dependency grammar. *Computational Linguistics*, 16(4), 234-236.

Cresti E. *et al.* (2002) The C-ORAL-ROM project. New methods for spoken language archives in a multilingual romance corpus. Actes *3rd International Conference on Language Resources and Evaluation*. *LREC'2002*. Las Palmas de Gran Canaria. Espagne. vol. I, 2-9.

Cunningham H. (2000) A definition and short history of language engineering. *Natural Language Engineering*. 5(1), 1-16.

Danieli M., Gerbino E. (1995) Metrics for evaluating dialogue strategies in a spoken langage system. Actes *AAAI Spring symposium on empirical methodes in discourse interpretation and generation*. Standford, CA. 34-39.

Debili F. (1977) Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage. Doctorat Paris VIII, Paris, France.

Deerwester S., Dumais S., Landauer T. Furnas G., Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*. 416(6), 391-407

DELIC (2002) Le corpus de référence de français parlé. 2èmes journées de la linguistique de corpus, Lorient, France, p. 41 (résumé).

Deligne S., Bimbot F. (1995) Language modeling by variable length sequences : theoretical formulation and evaluation of multigrams. Actes *International Conference on Acoustics, Speech and Signal Processing*, Actes *IEEE International Conference on Acoustics, Speech and Signal Processing*, *ICASSP'1995*, Detroit, MI. 172-175.

De Mori R. (1994) Apprentissage automatique pour l'interprétation sémantique. Actes des *XX° Journées d'Etudes sur la Parole*, *JEP'1994*. Trégastel, France. 11-19.

Devillers L., Maynard H., Paroubek P. (2002) Méthodologies d'évaluation du dialogue parlé : réflexions et expériences autour de la compréhension. *TAL*, vol. 43, n° 2, 155-184.

Dias G., Guillore S., BassanoJ.-C., Pereira Lopes J. C. (2000) Extractions automatiques d'unités lexicales complexes : un enjeu fondamental pour la recherche documentaire. *TAL*, vol. 41 n° 2. 447-472. Hermès, Paris, France

Dister A. (2002) Normalisation de corpus oraux retranscrits : jusqu'à quel point ? Actes des *2ème journées de la Linguistique de Corpus*, Lorient, France. p. 15 (résumé).

Dutoit T. (1997) An introduction to text-to-speech synthesis. Kluwer Academic Publ., Dordrecht, Pays-Bas.

Dutoit T. (2000) Synthèse de la parole à partir d'un texte, In Boite R., Bourlard H., Dutoit H., Hancq J. et Leich H. *Traitement de la parole*, Coll. Electricité, Presses Polytechniques et Universitaires Romandes, Lausanne, Suisse. 345-442.

Dybkjaer L., Bersen N.-O. (2000) Usability issues in spoken dialogue systems. *Natural Language Engineering*, 6 (3-4), 243-271.

Dymetman M. (1994), Quelques développements récents à la périphérie de la Traduction Automatique, actes *TALN'94*, Marseille, France, 24-30.

Ejerhed E. (1993) Nouveaux courants en analyse syntaxique, *TAL*, 34(1), 61-82.

Estival D *et al*. (1994) Survey of existing Test Suites, rapport de recherche du LRE 62-089 D-WP1, University of Essex, Royaume-Uni.

Fellbaum C. (1998). Wordnet, an electronic database. MIT Press, Cambridge, MA.

Ferret O., Grau B., Hurault-Plantet M., Illouz G., Jacquemin C. (2001) Utilisation des entités nommées et des variantes terminologiques dans un système question-réponse. Actes *TALN'2001*, Tours, France. 153-162.

Ferret O., Grau B., Hurault-Planter M., Illouz G., Jacquemin C. (2000) QALC : the question-answering system of LIMSI-CNRS. Actes *Text REtrieval Conference*, *TREC-9*, Gaithersburg, Maryland, USA. 235-244.

Fillmore C. J. (1968) The case for case In Bach E., Harms R. (Eds) Universals in Linguistic Theory. Holt et Rinehart and Winston Inc, New-York, Etats-Unis. 1-90.

Fisher-Lokou J., Guéguen N., 2001, Impact of a mediator, mutual representation of the negociators and decision making in a dyad : evaluation in the case of computer-mediated-communication, *Studia Psychologica*, 43(1), 13-21

Fluhr C. (2000) Indexation et recherche d'information textuelle. In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection I^{2}C. Hermès, Paris, France. 235-251.

FRACAS consortium. (1996). Using the framework. *Public deliverable of the Fracas project*. *LRE 62-051*, Deliverable D16 (chapitre 3).

Fraser N. (1997) Assessment of interactive systems. *In* Gibbon D., Moore R., Winski R. (Eds.). (1997) Handbook of standards and resources for spoken language systems. Mouton de Gruyter, Berlin, Allemagne. 564-615.

Fraser N., Gilbert G. (1991) Simulating speech systems. *Computer Speech and Language*, 5. 81-99.

Fraser N., Gilbert G. (1991) Effects of system voice quality on user utterances in speech dialogue systems. Actes *2nd European Conference on Speech Communication and Technology, Eurospeech'91*, Gènes, Italie, 57-60.

Gadet F. (1989) Le français ordinaire, Colin, Paris, France.

Gadet F. (1992) Le français populaire, PUF, Paris, France.

Gadet F. (Dir.) (1992) Hétérogénéité et variation : Labov, un bilan. *Langages*, 108, Larousse, Paris, France.

Gadet F. (1999) La variation diaphasique en syntaxe. In Barbiéris J.-M. (Ed.) Le français parlé : variété et discours. *Praxiling*, Université de Montpellier III. 211-228.

Gaiffe B., Romary L. et Pierrel J-M. (1991) Reference in a multimodal dialogue: towards a unified processing. Actes *2nd European Conference on Speech Communication and Technology*. *Eurospeech'91*, Gènes, Italie.

Gasiglia N. (2002) Vers un corpus thématisé de dialogues radiodiffusés : défense et illustration. Actes *2ème journées de la Linguistique de Corpus*, Lorient, France. p. 19 (résumé).

Gendner V., Illouz G., Jardino M., Monceaux L., Paroubek P., Robba I., Vilnat A. (2002) A protocol for evaluation analyzers of syntax (PEAS). Actes *3rd International Conference on Language Resources and Evaluation*, *LREC'2002*. Las Palmas de Gran Canaria, Espagne. 590-596.

Gibbon D., Moore R., Winski R. (Eds.). (1997) Handbook of standards and resources for spoken language systems. Mouton de Gruyter, Berlin, Allemagne.

Gillet J., Ward W. (1998) A language model combining trigrams and stochastic context-free grammars. Actes *5th International Conference on Spoken Language Processing*, *ICSLP'1998*, Sidney, Australie. 2319-2322.

Glandière M. (1983) SPARTE. *Education et informatique*, 17.

Glass J (1999) Challenges for spoken dialogue systems. Actes *IEEE ASRU Workshop*. Keystone, Colorado, Etats-Unis.

Godfrey J. J., Zampolli A. (1995) Language resources : overview. *In* Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. 441 :444.

Goulian J. (2000) Analyse linguistique détaillée pour la compréhension automatique de la parole spontanée, Actes *RECITAL'2000*, Lausanne, Suisse.

Goulian J. (2002) Stratégie d'analyse détaillée pour la compréhension automatique robuste de la parole. Thèse Université de Bretagne Sud, Vannes, France. 13 Décembre 2002.

Goulian J., Antoine J.-Y. (2001) Compréhension automatique de la parole combinant syntaxe locale et sémantique globale pour une CHM portant sur des tâches relativement complexes, Actes *TALN'2001,* Tours, France. 203-212.

Goulian J., Antoine J-Y., Poirier F. (2002) Compréhension automatique de la parole et TAL : une approche syntaxico-sémantique pour le traitement des inattendus structuraux du français. Actes *TALN'2002*, Nancy, France. 389-394.

Grosz B.J., Joshi A.K., Weinstein S. (1995) Centering : a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2), 203-225.

Guenthner F., Krüger-Thielmann K., Pasero R., Sabatier P. (1992) Communication aids for ALS patients. Actes *3rd International Conference on Computers for the handicapped persons*, *ICCHP'1992*, Vienne, Autriche. 303-307.

Guenthner F., Langer S., Krüger-Thielmann K., Pasero R., Richardet N., Sabatier P. (1992) KOMBE : communication aids for the handicaped. rapport technique 92-55, CIS, Universität München, Munich, Allemagne.

Gufstafson J., Larsson A., Carlson R., Hellman K. (1997) How do system questions influence lexical choices in user answers ? Actes *5th European Conference on Speech Communication and Technology*. *Eurospeech'97*, Rhodes, Grèce. 2275-2278.

Gufstafson J., Bell L. (2000) Speech technology on trial : experience from the August system. *Natural Language Engineering* , 6 (3-4). 273-286.

Guyomard M., Nerzic P., Siroux J. (1993) Plan, métaplans et dialogue. Actes de la *4ème école d'été sur les traitements des langues naturelles*. Lannion, France.

Guyomard M., Siroux J. (1987) Experimentation in the specification of an oral dialogue. In Nieman et al. (Eds.), *Recent Advances in Speech Understanding and Dialog Systems*, Springer Verlag, Berlin, Allemagne, vol. 46. 497-501.

Guyomard M., Siroux J., Cozannet A. (1990) Le rôle du dialogue pour la reconnaissance de parole. Le cas du système Pages Jaunes. Actes *XVIII° Journées d'Etudes sur la Parole*, *JEP'1990*. Montréal, Canada. 322-326.

Habert B., Nazarenko A., Salem A. (1997). Les linguistiques de corpus. Armand Colin, Paris, France.

Harabagiu S, Moldovan D. (2000) FALCON : boosting knowledge for answer engines. Actes *Text Retrieval Conference*, Gaithersburg, Maryland, USA.. 479-488.

Harabagiu S., Pasca M., Maiorano J. (2000) Experiments with open-domain textual question answering, actes *18th Conference on Computational Linguistics, COLING'2000*, Saarbrücken, Allemagne. 292-298.

Harman D. (1992) User-Friendly Systems Instead of User-Friendly Front-Ends. *JASIS* 43(2): 164-174

Harman D., Schäube P., Smeaton A. (1995) Document retrieval. *In* Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. 259-265.

Hauptman A., Rudnicky A. (1988) Talking to computers : an empirical investigation. *International Journal of Man-Machine Studies*, 28. 583-604.

Heeman P. A., Allen J. F. (1999) Speech repairs, intonational phrases and discourse markers : modeling speakers utterances in spoken dialogue. *Computational Linguistics*, 25(4), 527-573.

Henisz-Dostert B., Macdonald R., Zarechnak M. (Eds.) (1979) Machine Translation. Mouton De Gruyter, Berlin, Allemagne.

Hindle D. (1983) Deterministic parsing of syntactif non fluencies. Actes *21th Annual meeting of the Association for Computational Linguistics, ACL'1983*, MIT, Cambridge MS. 123-128.

Hindle D (1989) Acquiring disambiguisation rules from text. Actes *27th Annual meeting of the Association for Computational Linguistic ACL'1989*, Vancouver, Canada. 118-125.

Hirschman L. *et al*. : MACDOW Group (1992) Multi-Site Data Collection for a spoken language Corpus. *Actes DARPA Specch and Natural Language Workshop*. 7-14.

Hirschman L., Thomspson H. S. (1995) Overview of evaluation in speech and natural language processing. In Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon, Etats-Unis. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. 475-481.

Hirschman L. (1998) Language understanding evaluations : lessons learned from MUC and ATIS, Actes *2nd Conference on Language Resources and Evaluation*, *LREC'98*. Grenade, Espagne, 117-122.

Hobbs J. R. (1978) Resolving pronoun references. *Lingua*, 44, 331-338.

Holan T., Kubon, Oliva K., Plátek M. (2000) On complexity of word order, *Traitement Automatique des Langues*, *TAL*., 41(1), 273-300, Hermès, Paris.

Hone K.S., Graham R. (2000) Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, 6 (3-4). 287-303

Huang X., Acero A., Hon H-W. (2001) Spoken language processing : a guide to theory, algorithm and system development. Prentice Hall, Upper Saddle River, NJ.

Hudson R. (2000) Discontinuity. *Traitement Automatique des Langues*, *TAL*, 41(1), 15-56, Hermès,

Paris.

Ittycheriah A., Franz M., Zhu W.J., Ratnaparkhi A. (2000) IBM Statistical Answering System. Actes *Text Retrieval Conference*, *TREC-9*. Gaithersburg, Maryland, USA. 229-238.

Ide N., Macleod C. (2001). The American National Corpus : a standardized resource for American English. Actes *Corpus Linguistics'2001*, Lancaster, Royaume-Uni, 274 :280.

Ide N. et Véronis J. (Dir.) (1999) Selected papers from TEI10 : celebrating the 10th anniversary of the Text Encoding Initiative, *Computers and the Humanities*, 33(1-2), Kluwer Academic Publishers , Dordrecht, Pays-Bas.

Issar S., Ward W. (1993) CMU's robust spoken language understanging system. Actes *3rd European Conference on Speech Communication and Technology*, *Eurospeech'9*3, Berlin, Allemagne. 2147-2151.

Jacobson I. (1993) Le génie logiciel orienté objet : une approche fondée sur les cas d'utilisation. ACM Press, Addison-Wesley.

Jamoussi S., Smaïli K., Haton J.-P. (2003) Vers la compréhension automatique de la parole : extraction de concepts par réseaux bayésiens. Actes *TALN'2003*, Batz-sur-Mer, France. 165-174

JeanJean C. (1984) Les ratés c'est fa- fabuleux, *Linx*, 10, 171-177.

Jelinek F. (1991) Up from trigrams ! The struggle for improved language models. *Actes 2nd European Conference on Speech Communication and Technology*, *Eurospeech'1991*, Gènes, Italie. 1037-1040.

Jelinek F. (1976) Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64, 532-556.

Jelinek F. (1998) Statistical methods for speech recognition. Language, Speech and Communication. MIT Press, Cambridge, MA.

Joshi A., Weir D. et Vijay-Shanker K. (1991) The convergence of mildly context-sensitive formalisms, In Sells P., Shieber S., Wasow T. (Eds.) *Foundations issues in Natural Language Processing*, MIT Press, Cambridge, MA.

Kahane S. (2000) Extractions dans une grammaire de dépendance lexicalisée à bulles, *Traitement Automatique des Langues, TAL*, 41(1), Hermès, Paris, France. 211-244.

Kamp H., Reyle U. (1993), *From discourse to logic*. Kluwer Academic Publ, Amsterdam, Pays-Bas

Karsenty L. (2001) Quelles stratégies de transparence pour la gestion des erreurs de compréhension dans le dialogue vocal en langage naturel. Rapport Technique. IRIT, Toulouse, France.

Kennedy C., Boguraev B. (1996) Anaphora for everyone : pronominal anaphora resolution without a parser. Actes *16th Conference on Computational Linguistics, COLING-96*, Copenhague, Danemark. 113-118.

Kerbrat-Orecchioni C. (1999) L'oral dans l'interaction : une liberté surveillée. *Revue Française de Linguistique Appliquée*, *RFLA*, 4(2), 41-55.

King M. (1995) Human Factors and User Acceptability. In Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. 491-494.

Khudanpur S., Wu J. Maximum-entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling. *Computer Speech and Language*, 14(4), 283-332.

König E. (1996) Introduction to categorial grammars. rapport de recherche, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart Allemagne.

Koo M. W. *et al*. (1995) KT-STS : A speech translation system for hotel reservation and a continuous speech recognition system for speech translation, Actes *4th European Conference on Speech Communication and Technology*, *Eurospeech'95*, Madrid, Espagne, 1227:1231.

Koskiennnemi K. (1990) Finite state parsing and disambiguation. Actes *13th Conference on Computational Linguistics*, *COLING'90*. Helsinki, Finlande. 229-232.

van Kuppevelt, Heid U., Kamp H. (2000) Best practice in spoken language dialogue systems engineering. *Natural Language Engineering*, 6 (3-4). 205-212.

Kurdi M-Z (2002) A spoken language understanding approach which combines the parsing robustness with the interpretation deepness. Actes *International Conference on Artificial Intelligence*. *ICAI'2001*. Las Vegas, Etats-Unis.

Kurdi M.Z. (2003) Analyse linguistique robuste et profonde du langage oral spontané. Thèse Université Joseph Fourier, Grenoble, France.

Kurdi M. Z., Ahafhaf M. (2002) Toward an objective and generic method for spoken language understanding systems evaluation : an extension of the DCR method. Actes *3rd International Conference on Language Resources and Evaluation*, *LREC'2002*. Las Palmas de Gran Canaria, Espagne. 545-550.

Kwok K.L., Grunfeld L., Dinstl N., Chan M. (2000) TREC-9 Cross Language, Web and Question-Answering Track Experiments using PIRCS. Actes *Text Retrieval Conference*, *TREC-9*. Gaithersburg, Maryland, USA. 419-428.

Lamel L., Rosset S., Bennacef S., Bonneau-Maynard H., Devillers L., Gauvain J.-L. (1995) Developpement of spoken language corpora for travel information. Actes *4th European Conference on Speech Comunication and Technology*, *Eurospeech'95*, Madrid, Espagne. 1961-1964.

Langlais P. (2002) Ressources terminologiques et traduction probabiliste : premiers pas positif vers un système adaptatif. Actes *TALN'2002*, Nancy, France. 43-53.

Langlais P., Simard M. (2003) De la traduction probabiliste aux mémoires de traduction (ou l'inverse). Actes *TALN'2003*, Batz-sur-Mer, France. 195-204.

Lappin S., Leass H.J. (1994) An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4), 535-561.

Lecomte A. (1996) Grammaire et théorie de la preuve : une introduction. *TAL*, 37(2). 1-38.

Le Cunff C. (2002) De l'usage des corpus en didactique de l'oral : recherche et formation. Actes des *2ème journées de la Linguistique de Corpus*, Lorient, France. p. 25 (résumé).

Leech G., Garside R. (1991) Running a grammar factory : the production of syntactically analysed corpora or " treebanks ". In Johansson S., Strenström A.-B. (Eds.) *English computer corpora : selected papers and research guide*, Mouton de Gruyter, Berlin, Allemagne. 15-32.

Leech G., Garside R., Atwell E. (1983) The automatic grammatical tagging of the LOB corpus. *ICAME News*, 7, 13-33.

Leech G., Garside R., Bryant M. (1994). CLAWS4 : The tagging of the British National Corpus. Actes *14th International Conference on Computational Linguistics*, *COLING'1994*, Kyoto, Japon, 622 624.

Leech G. (1997) Introduction. *in* Garside R., Leech G., McEnery A. (Eds.) *Corpus annotation : linguistic information from computer text corpora*. Longman, London, UK. 1:18.

Lefèvre F., Gauvain J.-L., Lamel L. (2002) Développement d'une technique générique pour la reconnaissance de la parole indépendante de la tâche. Actes *XXIV° Journées d'Etudes sur la Parole*, *JEP'2002*, Nancy, France, 221-224.

Lehmann, D. Estival, Oepen S. (1996). TSNLP : des jeux de phrases-test pour l'évaluation d'applications dans le domaine du TAL, Actes *TALN 96*, Marseille, France. 97-103.

Léon J. (1992) De la traduction automatique à la linguistique computationnelle. Contribution à une chronologie des années 1959-1965. *TAL*, 1992(1-2), 25-44.

Letellier-Zarshenas S., Nicolas P., Goulian J., Antoine J.Y. (1999) Inattendus structurels et

communication orale finalisée : influence de la tâche et du contexte interactif, Actes *Journées Internationales de Linguistique Appliquée*, *JILA'99*, Nice, France. 176-179.

Levelt W. (1983) Monitoring and self-repair in speech, *Cognition*, 14. 41-104

Levin E., Pieraccini R. (1992) Chronus : the next generation. *Speech Communication*, 11, 283-288.

Levin E., Pieraccini R. (1995) Concept-based spontaneous speech understanding. Actes *4th European Conference on Speech Communication and Technology*, *Eurospeech'95*, Madrid, Espagne. 555-558.

Levin E. et Pieraccini R. (1997) A stochastic model of computer-human interaction for learning dialogue strategies. Actes *5th European Conference on Speech Communication and Technology*, *Eurospeech'97*. Rhodes, Grèce. 1883-1886.

Life A., Salter I., Temem J.N., Dartigues H., Guidon A., Rosset S., Bennacef S., Lamel L. (1996) Data collection for the MASK kiosk : Woz vs prototype system. Actes *4th International Conference on Spoken Language Processing*, *ICSLP'1996*, Philadelphie, PA, Etats-Unis. 1672-1675. Disponible sur la Toile : http://www.asel.udel.edu/icslp/cdrom/vol3/658/a658.pdf

Litman D. J. (1985) Plan recognition and discourse analysis : an integrated approach for understanding dialogues. Thèse de Doctorat, U. de Rochester, Etats-Unis .

Lopez P. (1999) Représenter et utiliser les contraintes de la langue oral à l'aide d'une grammaire lexicalisée d'arbres adjoints. Actes *TALN'99*, Cargèse, France. 445-450.

de Loupy C., Bellot P., El-Bèze M., Marteau P.-F. (1998) Query expansion and classification of retrieval documents. Actes *Text Retrieval Conference, TREC-7,* Gaithersburg, Maryland. 382-389.

Magnuson, T. (1995) Word Prediction as Linguistic Support for Individuals with Reading an Writing Difficulties. Actes *TIDE : The European context for assistive technology*. Paris, France. 316-319.

Marcus M., Santorini B., Marcinkiewicz M. (1993). Building a large annotated corpus of English : the Penn Treebank. *Computational Linguistics*, 19(2), 313-330.

Marshall I. (1983) Choice of grammatical word-class without global syntactic analysis : tagging words in the LOB corpus. Computers and the Humanities, 17, 139-150.

Martinie B. (2001) Remarques sur la syntaxe des énoncés réparés en français parlé. *Recherches sur le Français Parlé*, 16 (2001), 189-206.

Maurel, D. Rossi, N. Thibault, R. (2001) Handias : un système multilingue pour l'aide à la communication de personnes handicapées. Actes *TALN'2001*, conférence associée *Ingénierie des langues et handicap*, Tours, France. vol. 2, 203–212

Maynard H., Lefèvre F. (2002) Apprentissage d'un module stochastique de compréhension de la parole. Actes *XXIV° Journées d'Etudes sur la Parole, JEP'2002*, Nancy, France. 129-132.

McCoy, K. F., Demasco, P. (1995) Somme applications of natural language processing to the field of augmentative and alternative communication Actes *IJCAI'95 Workshop on Developing AI Applications for Disabled People*, Montreal, Canada. 97-112.

Méloni H. (1982), Etude et réalisation d'un système de reconnaissance automatique de la parole continue, Doctorat d'Etat, Université d'Aix-Marseille II, France.

Ménier G. (1995) Système en ligne de lecture d'écriture cursive manuscrite. Analyse continue des primitives et interprétation globale optimisée par algorithme génétique. Doctorat Université Rennes 1, Rennes, France.

Merialdo B. (1994) Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2), 155-172.

Minel J.-L., Nugier S., Piat G. (1997) Comment apprécier la qualité des "résumés" automatiques de textes ? Les exemples des protocoles FAN et MLUCE et leurs résultats sur SERAPHIN. Actes

1ères Journées Scientifiques et Techniques du réseau FRANCIL, *JST-FRANCIL'97*, Avignon, France. 227-232.

Minker W. (1995) An English version of the LIMSI L'ATIS System. Rapport technique LIMSI 95-12. Orsay, France.

Minker W. (1998). Evaluation methodologies for interactive speech systems. Actes *1st International Conference on Language Resource and Evaluation*, *LREC'98*, Grenade, Espagne, 199-206.

Minker W. (1999) Compréhension automatique de la parole. L'harmattan, Paris, France.

Minker W., Bennacef S. (1996) Compréhension et évaluation dans le domaine ATIS. Actes *XXI° Journées d'Etudes sur la Parole*, *JEP'1996*, Avignon, France. 415-419.

Minker W., Waibel A., Mariani J. (1999) *Stochastically based semantic analysis*, Kluwer Academic Publ., Dordrecht, Pays-Bas.

Mitkov R. (1998) Robust pronoun resolution with limited knowledge. Actes *36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, *ACL-COLING'98*, Montréal, Canada, 869-875.

Mitkov R., Boguraev B., Lappin S. (Eds.) (2001) Special issue on computer anaphora resolution. *Computational Linguistics*, 27(4).

Moeschler J. (1989) Modélisation du dialogue. Hermés, Paris, France.

Monbet V., Marteau P.-F. (2001), Continuous Space Discrete Time Markov Models for Multivariate Sea State Parameter Processes. Actes *11th International Offshore and Polar Engineering Conference & Exhibition.* Stavanger, Norvège.

Monceaux L., Robba I. (2002) Les analyseurs syntaxiques : atouts pour une analyse des questions. Actes *TALN'2002*, Nancy, France. 195-204.

Moorgat M. (1997) Categorial type logics. *In* van Benthem J., ter Meulen A. (Eds.) *Handbook of logic and language*. Elsevier Sciences, North-Holland, Amsterdam, Pays-Bas, 93-177.

Morel M.A. (Ed.) (1989) Analyse linguistique d'un corpus ; 2° corpus : centre d'information et d'orientation de l'Université Paris V. Publications de la Sorbonne Nouvelle, Paris, France.

Mueller C., Strube (2001) Annotating anaphoric and bridging relations with MMAX. Actes *2nd SIGdial workshop on discourse and dialogue*. Aalborg, Danemark. 90-95.

van Noord G., Bouma G., Koeling R., Nederhof M.J. (1999) Robust grammatical analysis for spoken dialogue systems. *Natural Language Enginerring*, 5(1).

Oerder M. & Aust H. (1994) A realtime prototype of an automatic inquiry system, Actes *International Conference on Spoken Language Processing*, *ICSLP'94*, Yokohama, Japon, 703-706.

Okada, H. Otsuka (1993) Incremental elaboration in generating spontaneous speech, actes *International Symposium on Spoken Dialogue*, *ISSD'93*, Tokyo, Japon, 49-52.

den Os E., Boves L., Lamel L., Baggia P. (1999). Overview of the ARISE project, Actes *6th European Conference on Speech Communication and Technology*, *Eurospeech'99*, Budapest, Hongrie. 1527-1530.

Ozkan H., Bisseret A., Caelen J. (1991) Analyse de dialogues finalisés dans une perspective communicationnelle. Actes des *3èmes journées sur l'ingénierie des Interface Homme - machine*, *IHM'91*, Dourdan, France, 175-186.

Pallet D., Fiscus J., Fisher W., Garofolo J., Lund B., Prysboski M. (1994) 1993 benchmark tests for the ARPA spoken language program. actes *1994 ARPA Human Language Technology workshop*. Morgan Kaufman, Princeton, NJ. 49-74.

Pallet D.S., Fiscus J.G. et al. (1995) 1994 benchmark tests for the ARPA spoken language program. *Actes 1995 ARPA workshop on spoken language technology*. Morgan Kaufman, Princeton, NJ. 5-36.

Paroubek P., Rajman M. (2000) Etiquetage morpho-syntaxique, In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection *IC2*, Hermès, Paris. 131-150.

Pérennou G. (1996) Compréhension du dialogue oral : le rôle du lexique dans l'approche par segments conceptuels. Actes de l'atelier *Lexique et Communication Parlée*, GDR-PRC Communication Homme - machine, Toulouse, France. 169-178.

Pettier J.C., Guyomard M. (2000) Action modeling in dialogue context. Actes *3th Internation Workshop on Human-Computer Conversation*. Bellagio, Italie. 136-141.

Pieraccini R., Levin E., Eckert W. (1998) Spoken Language Dialogue: architecture and algorithms. Actes *XXII° Journées d'Etudes sur la Parole*, *JEP'98*, Martigny, Suisse, 387-395.

Pieraccini R., Levin E. (1995) A spontaneous-speech understanding system for database query applications, ESCA Workshop on Spoken Dialogue Systems. Vigso, Danemark. 85-88.

Pierrel J-M. (1987), Dialogue oral homme - machine. Hermès, Paris, France.

Pierrel J-M. (1988) Dialogue homme - machine en language naturel écrit et oral. Actes *1ères journées nationales du PRC Communication Homme - machine*, Ec2 Editions, Paris, France. 152-182.

Pierrel J-M., Romary L. (2000) Dialogue homme - machine. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 331-350.

Pierrel J-M. (dir.) (2000) Ingénierie des langues. *Collection I²C*. Hemès. Paris.

Polifroni J., Seneff S., Glass J., Hazen T.J. (1998) Evaluation methodology for a telephone-based conversational system. Actes *1st International Conference on Language Resource and Evaluation*, *LREC'98*, Grenade, Espagne, 43-49

Pollard C., Sag I.(1994) Head-driven Phrase Structure Grammar. University of Chicago Press, Chicago, Michigan.

Prasad R., Sarkar A. (2000) Comparing test-suite based evaluation and corpus-based evaluation of a wode-coverage grammar for english. Actes *2nd International Conference on Language Resource and Evaluation, LREC'2000 Workshop on using evaluation within HLT programs : results and trands*, Athènes, Grèce, 7-12.

Privat R. (2000), Interrogation multimodale de consultation de serveurs d'informations : application aux personnes âgées, Actes des *1ères Rencontres Jeunes Chercheurs en IHM*, *RJC-IHM'2000*, Ile de Berder, France, pp. 127-130.

Rajman M., Han J. (1995) Prise en compte de contraintes syntaxiques dans le cadre d'un système de reconnaissance de la parole, Actes *TALN'1995*, Marseille, France. 97-106.

Rambow O., Joshi A. (1994) A formal look at dependency grammars and phrase-structure grammars with special considerations of word-order phenomena, *In* Wanner L. (ed.), *Current issues in Meaning-Text Theory*, Pinter, Londres, Royaume-Uni.

Rastier F. (1987) Sémantique interprétative. PUF, Paris, France.

Reithinger N. et Klesen M. (1997), Dialogue act classification using language models. Actes *5th European Conference on Speech Communication and Technology*, *Eurospeech'97*. Rhodes, Grèce. 2235-2238.

Richardet N. (1998) Composition de phrase*s* assistée - Un système d'aide à la communication pour handicapés. Thèse de doctorat, Université de la Méditerranée, Marseille, France.

Ricco X., Dutoit T. (2001) Vers un logiciel multilingue et gratuit pour l'aide aux personnes handicapées de la parole : HOOK (une interface du projet W). Actes *TALN'2001*, conférence associée *Ingénierie des Langues et Handicap*, Tours, France. vol. 2, 223-232.

Richard P., Gaucher P., Maurel D. (2000), Projet CNHL : Chambre Nomade pour Handicapés Lourds, Actes *Handicap'2000*, Paris, France, pp. 101-107.

Roche E., Schabes Y (1997) Finite state language processing, MIT Press, Cambridge, MA.

(chapitre Deterministic Part of Speech Tagging with Finite State Transducers. 205-239).

Romary L. (2000) Outils d'accès à des ressources linguistiques. In Pierrel J.M. (Dir.) *Ingénierie des langues*. Coll. IC2. Hermès, Paris, France. 193-212.

Rozenknop A., Chappelier J.-C., Rajman M. (2003) Apprentissage discriminant pour les grammaires à substitution d'arbres. actes TALN'2003, Batz-sur-Mer, France. 225-234.

Rossato S., Blanchon H., Besacier L. (2002) Evaluation du premier démonstrateur de traduction de parole dans le cadre du projet NESPOLE ! Actes *atelier thématique " Couplage de l'écrit avec l'oral, TALN'2002*, Nancy, France. Vol 2, 149-161.

Rosset S. (2000) Stratégies et gestionnaire de dialogue pour des systèmes d'interrogation de bases de données à reconnaissance vocale. Doctorat Université Paris XI, Orsay, France. publié comme rapport de recherche 2001-18 du LIMSI-CNRS, Orsay, France. septembre 2001.

Rosset S., Lamel L. (2001) Gestionnaire de dialogue pour un système de dialogue à recon-naissance vocale. Actes *TALN'01*, Tours, France. 385-390.

Roussalany A., Pierrel J.-M. (1992) Dialogue oral homme-machine en langage naturel : le projet DIAL, *Techniques et Sciences Informatiques*, 11 (2), 45:91.

Roussel D., Kurdi M.-Z., Caelen J. (1999) Normalisation des extragrammaticalités, supertagging et analyse partielle pour le traitement de la parole. Actes *TALN'1999*, atelier « *méthodes hybrides TALN / TALP pour le traitement robuste du langage* ». Cargèse, France.

Sabah G. (1994) Projet DALI, *rapport d'activité GDR-PRC CHM*, 71-88. Disponible sur la Toile à l'adresse : http://www-geod.imag.fr/pages_html/projets/DALI.htm

Sabah G. (1989) L' Intelligence Artificielle et le langage. Hermès, Paris. 2 volumes.

Sabah G. (1992) Collaboration des sources de connaissances dans un système de traitement automatique des langues : l'exemple de CARAMEL, *in* J. Caelen (ed.), *Cognition, perception et action en communication parlée*.

Sabah G., Vivier J., Vilnat A., Pierrel J-M., Romary L, Nicolle A. (1997) Machine, langue et dialogue. L'Harmattan, Paris, France.

Sabatier P. (1997) Evaluer les systèmes de compréhension de textes, actes *1ères Journées Scientifiques et Techniques du réseau Francil*, *JST-FRANCIL'97*, Avignon, France, 223-226.

Salton G. (1972) Experiments in automatic thesaurus construction for information retrieval. Actes congrès de l'*IFIP'1972*, Ljubljana, Slovénie.

Salton G. (1989) Automatic text processing. The transformation, analysis and retrieval of information by computer. Addison-Wesley, New-York.

Samuelson C., Voutilainen A. (1997) Comparing a linguistic and a stochastic tagger. Actes ACL-EACL'1997, Madrid, Espagne.

Schadle I. (2003) Sibylle : système linguistique d'aide à la communication pour les personnes handicapées. Doctorat Université de Bretagne Sud, Vannes, France. 18 décembre 2003. Rapport de recherche VALORIA-CORAIL-2003-03.

Schadle I., Antoine J-Y., Memmi D. (1999), *Connectionist language models for speech understanding : the problem of word order variations*, Eurospeech'99, Budapest, Hongrie.

Schadle I., Antoine J-Y., Le Pévédic B., Poirier F., (2002) SibyLettre, prédiction de lettres pour la communication augmentée, revue *RIHM*, Vol3, n°2, ISSN 1289-2963, Europia, Paris, France, 2002

Schukat-Talamazzini E.G., Hendrych R., Kompe R., Niemann H. (1995) Permugram language models, Actes *4th European Conference on Speech Communication and Technology*, *Eurospeech'95*, Madrid, Espagne, 1773-1776.

Schwartz R. , Nguyen L., Makhoul J. (1996). Muliple-pass search strategies. In Lee C.H., Soong F.K., Paliwal K.K. (Eds.) *Automatic speech and speaker recognition*. Kluwer Academic Publ.,

Dordrecht, Pays-Bas.

Schwartz R., Chow Y.-L. (1990) The N-best algorithm : an efficient and exact procedure for finding the N most likely sentence hypotheses. Actes *IEEE International Conference on Acoustics, Speech and Signal Processing*, I*CASSP'1990*., Albuquerque, NM. 81-84.

Seneff S. (1992) TINA: a natural language system for spoken language applications. *Computational Linguistics*, 18(1). 61-86.

Seneff S. (1992) Robust parsing for spoken language systems. Actes *International Conference on Acoustics, Speech and Signal*, *ICASSP'1992*. San Francisco, Etats-Unis. 189-192

Seneff S., Mc Candless M., Zue V. (1995) Integrating natural language into the word graph search for simultaneous speech recognition and understanding, actes *4th European Conference on Speech Communication and Technology* , *Eurospeech'95*, Madrid, Espagne, 1781-1784.

Shannon C. (1948) The mathematical theory of communication. *Bell System Technical Journal*, 27, 398-403.

Siroux J., Guyomard M., Jolly Y., Multon F., Remondeau C. (1995) Speech and tactile-based GEORAL system. Actes *3rd European Conference on Speech Communication and Technology* , *Eurospeech'95*, Madrid, Espagne, 1943-1946.

Siroux *et al*. (1997) Multimodal reference in GEORAL Tactile. Actes du workshop « *refering phenomena in a multimedia context and their computational treatments* », SIGMEDIA et ACL/EACL, Madrid, Espagne. 39-44.

Sleator D. D. K., Temperley D. (1991) Parsing English with a link grammar , *rapport de recherche CMU-CS-91-196*, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA.

Spérandio J.-C., Létang-Figeac C. (1986) Simulation expérimentale de dialogues oraux en communication homme - machine. Rapport final GRECO Communication Parlée. CNRS, Paris, France.

Spivey-Knowlton M. J. (1992) Another context effect in sentence processoing : implications for the principle of referential support, *14th Annual Conference of the Cognitive Science Society*, Bloomington, USA. 486-491.

Srihari S.N. et Srihari R.K. (1995). Written Language Input : overview. In Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. 71-76.

Su. K-Y. et al. (1992) A unified framework to incorporate speech and language information in spoken language processing. Actes *IEEE International Conference on Acoustics, Speech and Signal Processing*, *ICASPP'1992*, San Francisco, CA. vol. 1, 185-188.

Tesnière L. (1959) Elements de syntaxe structurale. Klincksiek, Paris, France.

Toulotte J.M., Baudel-Cantegrit B., Trehou G (1990) Acceleration method using a dictionnary acess in a BLISS communicator. Actes *1990 biennal Conference of the International Society for Augmentative and Alternative Communication*. Stockhom, Suède.

Turing A. M. (1950) Computing machinery and intelligence. *Mind*. 59, 236.

Vaillant, P. (1997) PVI : Système de traduction d'icônes en langue, Interaction entre modalités sémiotiques : de l'icône à la langue. Thèse de Doctorat, Université Paris-XI, Orsay, France.

Valli A. et Véronis J. (1999) Etiquetage grammatical de corpus de parole : problèmes et perspectives. *Revue Française de Linguistique Appliquée*, 4(2), 113-133.

Vanderveken D. (1994) A complete formulation of a simple logic of elementary illocutionary acts. In Tsohatzidis S. L. (Ed.) *Foundations of speech act theory : philosophical and linguistic perspectives*. Routldege. 99-131.

Vergnes J., Giguet E. (1998) Regards théoriques sur le tagging. Actes *TALN'1998*, Paris, France. 22-31.

Véronis J. (1997) Une action d'évaluation des systèmes d'alignement de textes multilingues, Actes *1ères Journées Scientifiques et Techniques du réseau FRANCIL*, *JST-Francil'97*, Avignon, France. 191-198

Véronis J. (2000) Alignement de corpus multilingues. In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection I²C. Hermès, Paris, France. 152-171.

Véronis J. (2000). Annotation automatique de corpus : panorama et état de la technique. In Pierrel J-M. (Dir.) *Ingénierie des langues*. Collection I²C. Hermès, Paris, France. 235 :250.

Véronis P., Langlais P. (2000) Evaluation of parralel text alignment systems : ARCADE. In Véronis J. (Ed.) Parallel Text Processing, Kluwer Academic Publ., Dordrecht, Pays-Bas.

Villaneau J., Antoine J.-Y., Ridoux O. (2001) Combining snytax et pragmatic knowledge for the understanding of spontaneous spoken utterances. Actes *4th International Conference on the Logical Aspects of Computational Linguistics*, *LACL'01,* Le Croisic, France. In *LNAI 2099*, Springer-Verlag, 279-295.

Villaneau J., Antoine J.-Y., Ridoux O. (2002) LOGUS, un système formel de compréhension du français parlé spontané. Actes TALN'2002, Nancy, France, vol. 1, 165-174

Villaneau J. (2003) Contribution au traitement syntaxico-pragmatique de la langue naturelle parlée: approche logique pour la compréhension de la parole. Doctorat l'Université de Bretagne Sud, Vannes, France. 6 décembre 2003. Rapport de recherche VALORIA-CORAIL-2003-02.

Villaseñor L., Caelen J. (1999) Une logique pour le dialogue coopératif homme-machine. Actes *2ème Colloque International sur l'Apprentissage Personne-Système*, *CAPS'98*, Caen, France. 53-62 .

Vincent - Le Pévédic B. (1997) Prédiction morphosyntaxique évolutive dans un système d'aide à la saisie de textes pour des personnes handicapées physiques : HandiAS. Thèse de doctorat de l'Université de Nantes, Nantes, France. 13 octobre 1997.

Waibel A. (1996) Interactive translation of conversationnal speech. *Computer*, 27(7). 41-48.

Waibel A., Lee K. (Eds) (1990) Readings in speech recognition. Morgan Kaufman, Princeton, NJ.

Walker M., Kamm C., Boland J. (2000) Developping and testing general models of spoken dialogue system performance, Actes *2nd International Conference on Language Resources and Evaluation*, *LREC'2000*, Athènes, Grèce. 189-196

Walker M., Kamm C., Litman D. (2000) Towards developping general models of usability with PARADISE. Natural Language Engineering, 6 (3-4), 363-377.

Walker M., Passonneau R., Boland J. (2001) Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialog Systems, Actes *ACL/EACL'2001*. Toulouse, France.

Wang Y., Mahakan M., Huang X. (2000) A unified context-free grammar and N-gram model for spoken language processing. Actes *International Conference on Acoustics, Speech and Signal Processing*, *ICASSP'2000*, Istambul, Turquie, 1639-1642.

Young S., Ward W. (1993) Semantic and pragmatically based recognition of spontaneous speech. Actes *3rd European Conference on Speech Communication and Technology*, *Eurospeech'9*3, Berlin, Allemagne. 2244-2247.

Zechner K. (1998) Automatic construction of frame representations for spontaneous speech in unrestricted domains. Actes *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, *COLING-ACL'1998*. Montréal, Canada. 1448-1452.

Zeiliger J., Caelen J., Antoine J.-Y. (1997) Vers une méthodologie d'évaluation qualitative des systèmes de compréhension et de dialogue oral homme - machine. Actes 1ères Journées Scientifiques et Techniques du réseau FRANCIL, JST-FRANCIL'97, Avignon, France. 437-446 ; Texte repris dans : Chibout K., Mariani J., Masson N., Néel F. (Dir.) (2000) Ressources et évaluations en ingénierie des langues. De Boeck Université, Duculot, Bruxelle, Belgique. 437-461.

Zitouni I. (2000) Modélisation du langage pour les systèmes de reconnaissance de la parole destinés aux grands vocabulaires. Doctorat de l'Université Nancy I, France.

Zweigenbaum P., Garbar N. (2002) *Accentuation de mots inconnus : application au thesaurus biomédical MeSH*. Actes TALN'2002, Nancy, France. 53-62.

Annexes A : Liste de publications

PUBLICATIONS — 1996 / 2003

Cette liste regroupe l'ensemble de mes publications postérieurs à ma nomination aux fonctions de maître de conférences de l'Université de Bretagne Sud. Cette liste a été arrêtée au 01/07/2003.

Revues avec comité scientifique

Antoine J.-Y, Vigouroux N. (Eds.) (2003), numéro spécial RJC-IHM'2000, *Revue d'Interaction Homme-Machine, RIHM*, vol. 4, Europia, Paris, France (à paraître).

Schadle I., Antoine J-Y., Le Pévédic B., Poirier F., (2002) SybiLettre, prédiction de lettres pour la communication augmentée, *Revue d'Interaction Homme-Machine, RIHM*, Vol 3, n°2, ISSN 1289-2963, Europia, Paris, France, 2002.

Antoine J.-Y, Goulian J. (2001), Étude des phénomènes d'extraction en français parlé sur deux corpus de dialogue oral finalisé : application à la CHM orale, *TAL*, 42(2), Hermès, 413-440.

Antoine J-Y., Caelen J. (1999), Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR (Demande, Contrôle, Résultat), *Langues*, 2 (2), 130-139.

Chapitres dans ouvrage

Zeiliger J., Antoine J-Y., Caelen J., (2000), La méthodologie DQR d'évaluation qualitative des systèmes de dialogue oral homme-machine, in J. Mariani et al. (Ed.), *Ressources et Evaluation en Ingénierie de la Langue*, coll. Universités francophones, série Actualité scientifique, AUF et De Boeck Université, Paris, 437-450.

Congrès internationaux avec comité scientifique

Goulian J, Antoine J.-Y., Poirier F. (2003) How NLP techniques can improve speech understanding : ROMUS, a robust chunk based message understanding system using link grammars. *8th European Conference on Speech Communication and Technology*, *Eurospeech'2003*, Genève.

Nicolas P., Letellier-Zarshenas S., Schadle I., Antoine J.-Y., Caelen Jean (2002) Towards a large corpus of spoken dialogue in French that will be freely available: the "*Parole Publique*" project and its first realisations. actes *3rd International Conference on Language Resources & Evaluation*, *LREC'2002*, Las Palmas de Gran Canaria, Espagne.

Antoine J.-Y., Bousquet-Vernhettes C., Goulian J., Kurdi M. Z., Rosset S., Vigouroux N., Villaneau J. (2002) Predictive and objective evaluation of speech understanding: the "challenge" evaluation campaign of the I3 speech workgroup of the French CNRS. actes *3rd International Conference on Language Resources & Evaluation*, *LREC'2002*, Las Palmas de Gran Canaria, Espagne.

Villaneau J., Antoine J-Y., Ridoux O., (2001)., *Combining syntax and pragmatic knowledge for the*

understanding of spontaneous spoken sentences , actes *Logical Aspects of Computational Linguistics*, *LACL'2001*, Le Croisic, France, publié in LNAI 2099, Springer, 279-295.

Antoine J-Y., Goulian J., (2001), *Word order variations and spoken man-machine dialogue in French : a corpus analysis on the ATIS domain*, Corpus Linguistics'2001, Lancaster, GB.

Antoine J-Y., Siroux J., Caelen J., Villaneau J., Goulian J., Ahafhaf M. (2000), *Obtaining predictive results with an objective evaluation of spoken dialogue systems : experiments with the DCR assessment paradigm*, LREC'2000, Athènes, Grèce.

Schadle I., Antoine J-Y., Memmi D. (1999), *Connectionist language models for speech understanding : the problem of word order variations*, Eurospeech'99, Budapest, Hongrie.

Antoine J.Y., Zeiliger J., Caelen J., (1998) *DQR Test suites for a qualitative evaluation of spoken dialogue systems : from speech understanding to dialogue strategy*, 1st International Conference on Language Ressources and Evaluation, LREC'98, Granada, Spain.

Antoine J.-Y. (1996), *Parsing spoken language without syntax: a microsemantic approach*, 16th International Conference on Computational Linguistics, COLING'96, Copenhague, Danemark.

Antoine J.-Y. (1996), *Spontaneous speech and natural language processing. ALPES : a robust semantic-led parser*, 4th International Conference on Spoken Language Processing, ICSLP'96, Philadelphie, PA, USA.

Workshops internationaux avec actes et comité scientifique

Antoine J.-Y., Zeiliger J., Caelen J., (1997) RQA methodology : towards a qualitative evaluation of speech understanding and spoken dialog, *SALT Workshop on Evaluation*, Sheffield U., Royaume-Uni.

Congrès nationaux avec comité scientifique

Antoine J-Y, Goulian J, Villaneau J (2003) Quand le TAL robuste s'attaque au langage parlé : analyse incrémentale pour la compréhension de la parole spontanée, actes *TALN'2003*, Batz-sur-Mer (à paraître)

Antoine J.-Y., Letellier-Zarshenas S., Schadle I., (2002) Le projet PAROLE PUBLIQUE de constitution d'un large corpus francophone de dialogue oral : réalisations et perspectives, actes *CORPLING'2002*, Lorient, France, Presses Universitaires de Rennes.

Antoine J.-Y., Letellier-Zarshenas S., Schadle I., Nicolas P., Caelen J. (2002) Corpus OTG et ECOLE_MASSY :
vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement, actes *TALN'2002*, Nancy, France.

Goulian J., Antoine J.-Y., Poirier F. (2002) Compréhension automatique de la parole et TAL : une approche syntaxico-sémantique pour le traitement des inattendus structuraux du français parlé actes *TALN'2002*, Nancy, France.

Villaneau J., Antoine J.-Y., Ridoux O. (2002) LOGUS : un système formel de compréhension du français parlé spontané : présentation et évaluation. actes *TALN'2002*, Nancy, France, 165-174.

Antoine J.-Y., Letellier-Zarshenas S., Schadle I., Nicolas P. (2002) Le projet PAROLE PUBLIQUE de constitution d'un large corpus francophone de dialogue oral : réalisations et perspectives, actes *2ème journées de la Linguistique de Corpus*, Lorient, France.

Goulian I., Antoine J.-Y (2001) Compréhension automatique de la parole combinant syntaxe locale et sémantique globale pour une CHM portant sur des tâches relativement complexes, Actes. *TALN'2001*, Tours, France, pp. 203-212

I. Schadle, Le Pévédic B., Antoine J.-Y., Poirier F. (2001), Prédictions de lettres pour l'aide à la saisie de texte, actes *JIM'2001*, Metz, France.

Antoine J-Y., Goulian J., (1999), *Le français est-il une langue à ordre variable* ?, Journées Internationales de Linguistique Appliquée, JILA'99, Nice, France.

Letellier-Zarshenas S., Nicolas P., Goulian J., Antoine J.Y. (1999) *Inattendus structurels et communication orale finalisée : influence de la tâche et du contexte interactif*, Journées Internationales de Linguistique Appliquée, JILA'99, Nice, France.

Zeiliger J., Caelen J., Antoine J.-Y., (1997), Vers une Méthodologie d'Evaluation Qualitative des Systèmes de Compréhension et de Dialogue Oral Homme-machine, *Actes Journées Scientifiques et Techniques FRANCIL*, Avignon, France.

Workshops nationaux avec actes et comité scientifique

Antoine J.-Y, Le Pévédic B. (2001), Ingénierie des Langues et Handicap, actes *TALN'2001*, atelier thématique « Handicap et Ingénierie Linguistique », Tours, France.

Schadle I., Le Pévédic B., Antoine J.-Y, Poirier F (2001), Sybillette : système de prédiction de lettre pour l'aide à la saisie de texte, actes *TALN'2001*, atelier thématique « Handicap et Ingénierie Linguistique », Tours, France, vol 2., p. 233-242

Antoine J.-Y., Genthial D. (1999), *Méthodes hybrides issues du TALN et du TAL Parlé : états des lieux et perspectives*, TALN'1999, atelier thématique « Méthodes hybrides pour le TAL robuste », Cargèse, France, 1-17.

Communications sans actes

Antoine J.-Y., Goulian J., Letellier-Zarshenas S. (2002), Corpus de dialogue oral pour la Communication Homme-Machine : quelques enseignements en linguistique et en Traitement Automatique des Langues Naturelles, journées d'étude de l'ATALA, Paris, France.

Colineau N., Rouibah A., Antoine J.-Y. (1997), Recueil et exploitation de grands corpus : apports de la Psychologie Expérimentale et de la Communication Homme - Machine, colloque de l'Association Française de Linguistique Appliquée, AFLA, Paris, France.

Rapports techniques

Antoine J.-Y. (2002) Corpus OTG : présentation générale. Rapport de recherche VALORIA-EQUIPAGE-LN-2002-2. http://www.univ-ubs.fr/valoria/antoine/parole_publique/OTG/Pres_OTG.pdf.

Antoine J.-Y. (2002) Corpus Ecole Massy : présentation générale. Rapport de recherche VALORIA-EQUIPAGE-LN-2002-1. http://www.univ-ubs.fr/valoria/antoine/parole_publique/Massy/Pres_Massy.pdf.

Antoine J.-Y. (2001) Méthodologie d'évaluation par défi. Rapport de recherche VALORIA-EQUIPAGE-LN-2001-1. http://www.univ-ubs.fr/valoria/antoine/gdri3/eval_defi.html.

Antoine J-Y. (1998), Corpus pilote OTG : conventions de transcription orthographique, Rapport de recherche VALORIA-EQUIPAGE-LN-1998-01.

PUBLICATIONS DES ETUDIANTS ENCADRES — 1996 / 2003

Cette liste regroupe les publications personnelles — i.e. où je ne suis pas co-auteur — des étudiants que j'ai encadré depuis mon arrivée à l'Université de Bretagne Sud. Cette liste a été arrêtée au 01/07/2003. Les DEA de J. Foulon, F. Lamie et I. Randria donneront lieu à soutenance publique en septembre 2003.

Mémoires de thèse

Goulian J. (2002) Stratégie d'analyse détaillée pour la compréhension automatique robuste de la parole. Thèse Université de Bretagne Sud, Vannes, France. 13 Décembre 2002. Rapport de recherche VALORIA-CORAIL-2002-03.

Villaneau J. (2003) Contribution au traitement syntaxico-pragmatique de la langue naturelle parlée: approche logique pour la compréhension de la parole. Doctorat l'Université de Bretagne Sud, Vannes, France. 6 décembre 2003. Rapport de recherche VALORIA-CORAIL-2003-02.

Schadle I. (2003) Sibylle : système linguistique d'aide à la communication pour les personnes handicapées. Doctorat Université de Bretagne Sud, Vannes, France. 18 décembre 2003. Rapport de recherche VALORIA-CORAIL-2003-03.

Mémoires de DEA

Goulian J. (1998) Analyse Robuste du français parlé. DEA Sciences Cognitives, INPG, Grenoble, France. Juin 1998.

Schadle I. (1998) Analyse flexible du français parlé par réseaux récurrents. DEA Sciences Cognitives, INPG, Grenoble, France.

Derouard L. (1997) Traitement robuste de la parole spontanée par réseaux récurrents. DEA Sciences Cognitives, INPG, Grenoble, France.

Communications personnelles

Goulian J. (2000) *Analyse linguistique détaillée pour la compréhension automatique de la parole spontanée*. Actes RECITAL'2000. Lausanne, Suisse.

Villaneau J. (2000) *Un système basé sur les types logiques pour la compréhension de la parole*. Actes RECITAL'2000. Lausanne, Suisse.

PUBLICATIONS — 1992 / 1996

Cette liste regroupe l'ensemble de mes publications antérieurs à ma nomination aux fonctions de maître de conférences.

Mémoire de thèse

Antoine J.-Y. (1994), *Coopération syntaxe-sémantique pour la compréhension automatique de la parole spontanée*, Doctorat d'université spécialité signal-image-parole, INPG, Grenoble, France.

Chapitres dans ouvrage

Antoine J.-Y. (1995), *Conception de dessins et Communication Homme-Machine : améliorer l'interaction orale au niveau linguistique*, in K. Zreik, J. Caelen (ed.), *Communication en conception*, EuropIA éditions.

Congrès internationaux avec comité scientifique

Antoine J.-Y. (1995), Caelen J., Caillaud B., *The multi-agent paradigm as a computing model of holistic-analytic vicariancy. MICRO : an illustration in automatic spoken language understanding*, 1st European Conference on Cognitive Science, ECCS'95, Saint-Malo, France, Avril 1995.

Antoine J.-Y. (1994), Caillaud B., Caelen J., *Automatic Adaptative Understanding of Spoken Language by Cooperation of Syntactic Parsing and Semantic Priming*, International Conference on Spoken Language Processing, 3rd International Conference on Spoken Language Processing, ICSLP'94, Yokohama, Japon.

Antoine J.-Y., Caillaud B., Caelen J. (1993), *Syntax-semantics cooperation in MICRO, a multi-agents speech understanding system.*, 3rd European Conference on Speech Communication and Technology, EUROSPEECH'93, Berlin.

Congrès nationaux avec comité scientifique

Antoine J.-Y. (1996), *ALPES, un modèle microsémantique robuste pour l'analyse du langage parlé*, Rencontres des Etudiants-Chercheurs en Informatique pour le Traitement Automatique de la Langue, RECITAL'96, Gif-sur-Yvette, France.

Antoine J.-Y., Caelen J. (1996), *Améliorer la reconnaissance de la parole par l'intégration de contraintes linguistiques robustes : le modèle microsémantique ALPES*, 21° Journées d'Etudes de la Parole, JEP'96, Avignon, France.

Antoine J.-Y. (1995), *Compréhension automatique de la parole spontanée et dialogue oral personne - système : vers une analyse linguistique sans syntaxe?*, 1° Journées Jeunes Chercheurs en Parole, SFA-GFCP, ENST, Paris, France.

Antoine J.-Y. (1995), *Conception de dessin et interaction orale : les niveaux linguistiques*, 4° Table Ronde Francophone sur la Conception ("Aspects Communicatifs en Conception"), 01 Design'95, Autrans, France.

Antoine J.-Y., Caillaud B. (1994), *Vicariance holistique-analytique en compréhension de la parole*,

1° colloque jeunes chercheurs en sciences cognitives de l'ARC, La Motte d'Aveillans, France.

Caillaud B., Antoine J.-Y., Caelen J., Caelen-Haumont G. (1994), *MICRO, un système multi-agents pour la compréhension de la parole*, 9° congrès Reconnaissance des Formes et Intelligence Artificielle, RFIA'94, Paris, France.

Antoine J.-Y. et al. (Pôle PLEAID) (1992), *Vers une taxonomie du vocabulaire pour les systèmes multi-agents*, Journées du PRC-IA sur les Systèmes Multi-Agents, Nancy, France.

Caillaud B., Antoine J.-Y., Caelen J. (1992), *Expertise sur le contrôle dans un système multi-agent de compréhension de la parole,* Journées PRC-IA sur les Systèmes-Multi-Agents, Nancy, France.

Communications sans comité scientifique

Antoine J.-Y. (1995), compte-rendu de la table ronde *Parole et linguistique*, 1° Journées Jeunes Chercheurs en Parole, SFA-GFCP, ENST, Paris, France.

Caelen J., Caillaud B., Antoine J.-Y. (1994), *Projet MICRO : Modélisation Informatique de la Cognition en Reconnaissance de l'Oral*, journées du PRC-CHM (Communication Homme-Machine) sur la Reconnaissance automatique de la parole, Nancy, France.

Antoine J.-Y., Caillaud B., Caelen J. (1994), *MICRO*, 2° journées Francophones en Intelligence Artificielle Distribuée et Systèmes Multi-Agents (vol. de présentation des démonstrations), Voiron, France.

Rapports techniques

Antoine J.-Y. (1996), *Evaluation de la compréhension dans le dialogue oral spontané : structures sémantiques élaborées par l'analyseur ALPES*, rapport technique AUPELF-UREF, réf. S1/5.4/ARC ILOR.B2t/2879.

Annexes B : résumés des activités de recherche

ANTOINE Jean-Yves

Laboratoire VALORIA (EA 2593) - Université de Bretagne Sud
IUP Vannes, rue Yves Mainguy, 56 000 Vannes
Mél : Jean-Yves.Antoine@univ-ubs.fr — Toile : http://www.univ-ubs.fr/valoria/antoine

SITUATION STATUTAIRE

Maître de Conférences en informatique (27° section CNU)

- **Nomination** 1er Septembre 1996, en qualité de MCF 2° classe
- **Titularisation** 1er Septembre 1997
- **Dernière Promotion** 1er Septembre 2000 à la 1ère classe MCF (classe unique depuis 2001)

Titulaire d'une Prime d'Encadrement Doctoral et de Recherche (PEDR) depuis le septembre 1999

ACTIVITES DE RECHERCHE ENTRE 1999 et 2003

1. ANIMATION ET ENCADREMENTS SCIENTIFIQUES

Direction du groupe de recherche CORAIL au sein du laboratoire VALORIA (1996-...)

L'équipe EQUIPAGE du laboratoire VALORIA, regroupe des thématiques de recherche relevant de l'Interaction Homme-Machine, de l'Apprentissage et de l'Ingénierie Linguistique. J'y assure depuis mon intégration au VALORIA la **direction scientifique** du groupe de recherche en Ingénierie Linguistique appelé CORAIL depuis 2001. Ce groupe comprend à l'heure actuelle : **3 Maîtres de Conférences**, **2 ATER** et **2 Doctorants**.

Au sein de ce groupe, trois axes de recherche ont été définis sous ma responsabilité : Communication Homme-Machine à composante langagière; traitements linguistiques et handicap; ressources linguistiques et évaluation des systèmes. Cette structuration découle d'une politique scientifique orientée vers la prise en compte des usages linguistiques réels dans les applications relevant des technologies langagières.

Direction du groupe de travail « Compréhension robuste » du GDR-I3 du CNRS (1998-...)

Je dirige depuis 1998 le groupe de travail « Compréhension de la parole » (GT 5.5) du GDR-I3 (Information-Interaction-Intelligence) du CNRS. Centré sur l'évaluation des systèmes de compréhension de parole, ce groupe de travail regroupe cinq laboratoires français : CLIPS-IMAG (Grenoble), IRIT (Toulouse), LIMSI-CNRS (Orsay), LIA (Avignon), LORIA (Nancy) et VALORIA.

Je suis par ailleurs **membre du comité de direction du GDR-I3** depuis 2002.

Rédacteur en chef *In Cognito* (1995-2001) et *Cahiers romans de sciences cognitives* (2002-...)

Initialement francophone, *In Cognito* est depuis 2002 une revue internationale en sciences cognitives publiée en quatre langues romanes (espagnol, français, italien, portugais) sous le nom des *Cahiers Romans de Sciences Cognitives*. Quelques données pour résumer cette revue (site WWW : http://www.in-cognito.net) :

- *disciplines* — Intelligence Artificielle et ses applications (vision, robotique, traitement d'images, TALN...), Interaction Homme-Machine, Sciences du langage, Psychologie cognitive, Neurosciences cognitives, Sociologie cognitive, Sciences de l'éducation, Philosophie.
- *diffusion* — essentiellement Europe, Amérique du Nord et Latine.
- *sélectivité* — Taux moyen de rejet : 65 %

Formation doctorale

Je suis coordonnateur pédagogique du DEA Information (mention Interaction Homme-Machine) de l'Université de Bretagne Sud et de l'ENST Bretagne (2002-2003).

Encadrements de thèse

- *J. Goulian* — Doctorat de l'U. de Bretagne Sud soutenu le 13 décembre 2002 — Directeur de thèse : F. Poirier (VALORIA, UBS) — Encadrant scientifique : J.-Y. Antoine.
- *J. Villaneau* — Doctorat de l'U. de Bretagne Sud — Directeur de thèse : O. Ridoux (IRISA, U. Rennes 1) — Co-encadrant : J.-Y. Antoine. Soutenance : novembre 2003.
- *I. Schadle* — Doctorat de l'U. de Bretagne Sud — Directeur de thèse : F. Poirier (VALORIA, UBS) — Co-encadrant : J.-Y. Antoine. Soutenance prévue pour fin 2003.
- *V. Bralé* — Doctorat en informatique de l'U. de Bretagne Sud — Directeur de thèse : I. Kanellos (ENST Bretagne) — Co-encadrant : J.-Y. Antoine.

Encadrements de DEA

- *J. Foulon* — DEA Informatique U. de Bretagne Sud (2003) — Encadrment J-Y. Antoine.
- *I. Randria* — DEA Informatique U. de Bretagne Sud (2003) — Encadrement J.Y. Antoine.
- *F. Lamie* — DEA Informatique U. de Bretagne Sud (2003) — Co-encadrement J-Y. Antoine et P.-Y. Nicolas (CCI Brest)
- *J. Goulian* — DEA Sciences Cognitives INPG (1998) — Encadrement J-Y. Antoine et D. Genthial (CLIPS-IMAG, Grenoble).
- *I. Schadle* — DEA Sciences Cognitives INPG (1998) — Encadrement J.Y. Antoine et D. Memmi (LEIBNIZ-IMAG).
- *L. Derouard* — DEA Sciences Cognitives INPG (1997) — Encadrement J.Y. Antoine et D. Memmi (LEIBNIZ-IMAG).

2. COLLABORATIONS

Relations avec la recherche académique

- **GT GDR-I3 « Compréhension de la parole » du CNRS** (1998-…) — Collaboration présentée en paragraphe 1.
- **Projets TECHNOLANGUE** (2002-2005)

Le groupe CORAIL participe sous ma direction à trois projets dans le cadre de l'appel d'offre TECHNOLANGUE du Ministère de le Recherche :

- Projet MEDIA-EVALDA sur l'évaluation des systèmes de compréhension de parole. Ce projet est la continuation des activités du GT 5.5. du GDR-I3 que je dirige. Participants : ELDA, VECSYS, IRIT, LIA, LIMSI, LORIA, VALORIA, TELIP, France Telecom R&D.

- Projet EASY-EVALDA sur l'évaluations des analyseurs syntaxiques. Notre groupe de recherche participera à cette campagne d'évaluation avec deux systèmes. Participants : ELDA, LIMSI, ATILF, LLF, DIAM, DELIC, GREYC-CNRS, LORIA, LPL, Synapse, Systal, Rank Xerox RC, CEA, LIA-EPFL, France Telecom R&D, Tagmatica, LATL, Talana, VALORIA.
- Projet OURAL-AGILE sur la constitution d'outils de base pour l'ingénierie des langues. L'intervention de notre groupe est centrée sur la fourniture de ressources linguistiques (corpus de dialogue oral). Participants : SINEQUA, LIP6, VALORIA, LPE, SILEX, LIMSI, LIA.

- **Commission de normalisation RNIL-AFNOR** (2002-…)

Je participe depuis 2002 aux activités de la commission de normalisation RNIL (Ressources Normalisées pour l'Ingénierie des Langues) de l'AFNOR (Association Française de Normalisation). Ce comité, qui regroupe des experts de l'ensemble de la communauté française en ingénierie des langues, tient lieu de représentant français au sous-comité TC37 / SC 4 de l'ISO consacré à la normalisation des ressources et traitements linguistiques.

- **Action ASILA du CNRS** (2002-2003)

Cette action, regroupant de nombreux partenaires et dirigée par le LORIA (Nancy), concerne l'étude du dialogue oral et écrit. Notre équipe participe à ce projet en qualité de fournisseur de ressources linguistiques orales.

- **Action de recherche concertée « Dialogue Oral » de l'AUPELF-UREF** (ARC-ILOR B2; 1996-2001)

Cette ARC, qui concernait l'évaluation des systèmes de dialogue oral, a réuni les participants suivants : LIMSI (Orsay), CLIPS-IMAG (Grenoble), IRIT (Toulouse), IRISA-Cordial (LLI Lannion) et VALORIA. Dans le cadre de cette action, notre groupe a réalisé un corpus de dialogue oral qui est désormais distribué librement sur le site de notre projet PAROLE PUBLIQUE (http://www.univ-ubs.fr/valoria/parole_publique)

- **Co-encadrements doctoraux**

CLIPS-IMAG	DEA Jérôme Goulian	(1998)
LEIBNIZ-IMAG	DEA Laurent Derouard et Igor Schadle	(1997-1998)
IRISA	Doctorat de Jeanne Villaneau	(1999-2003)
ENST-Bretagne	Doctorat de Véronique Bralé	(2002-2005)

Relations avec le monde industriel

Collaboration KERPAPE (2000-...) — A l'occasion du doctorat d'Igor Schadle, une collaboration s'est mise en place avec le centre de rééducation et réadaptation fonctionnelle de Kerpape, géré par la Mutualité du Morbihan. Cette collaboration est centrée autour de la réalisation et la validation in situ du systèmes d'aide linguistique à la saisie pour personnes handicapées réalisé par Igor Schadle.

Projets TECHNOLANGUE (2002-2005) — Dans le cadre des projets TECHNOLANGUES auxquels participe notre groupe de recherche, on peut citer les collaborations avec les partenaires industriels suivants : France Telecom R&D, SINEQUA, TELIP, VECSYS, Synapse, Systal, Rank Xerox RC, CEA, Tagmatica.

Co-encadrements

France Télécom R&D	Doctorat de Véronique Bralé	(2002-2005)
CCI - aéroport de Brest	DEA de Frédéric Lamie	(2003)

3. RAYONNEMENT SCIENTIFIQUE

Participation à des jurys de thèse

- Caroline Bousquet-Vernhettes (U. Paul Sabatier, Toulouse) septembre 2002, examinateur
- Jérome Goulian (U. Bretagne Sud, Vannes) décembre 2002, co-directeur
- Mohamed-Zakaria Kurdi (U. Joseph Fourier, Grenoble) avril 2003, examinateur

Instances représentatives

- Membre du comité directeur du GDR-I3 (Information – Intelligence - Interaction) du CNRS.
- Membre du conseil d'administration de l'ATALA (Association pour le Traitement Automatique des LAngues)

Comités scientifiques / comités de programme / comités de lecture

- Rédacteur en chef des revues *In Cognito* (1995-2001) et *Cahiers Romans de Sciences Cognitives* (2002 -...)

- *European Journal of Operational Research (EJOR) :* comité de lecture du numéro thématique « Human Cognitive Processes » (2000)
- *Revue d'Intelligence Artificielle* : comité de lecture du numéro thématique « Recherche d'information » (2003)
- *Traitement Automatique des Langues (TAL)* : comité de lecture du numéro thématique « Dialogue » (2002)
- *Revue d'Interaction Homme-Machine* : relecteur occasionnel (2002-2003)

- 14th Euro-conférence *Human Centered Processes 2003*, HCP'2003 (Luxembourg, 2003) : comité de programme

- RECITAL'2004 (Fès, Maroc, 2004) : comité de lecture
- TALN'2003 (Batz-sur-Mer, 2003) : comité de programme
- 4ème Colloque Jeunes Chercheurs en Sciences Cognitives, CJC'4 (Lyon, 2001) : comité de programme
- Atelier « TALN et Handicap », TALN'2001 (Tours, 2001) : comité scientifique
- 1ères Rencontres Jeunes Chercheurs en Interaction Homme-Machine (Baden, 2000) : comité de programme
- Atelier « Méthodes hybrides pour le TALN/TALP », TALN'1999 (Cargèse, 1999) : comité scientifique

- Comité scientifique du programme d'évaluation MEDIA-EVALDA.

Organisation de conférences

- Président du comité d'organisation des Journées Jeunes Chercheurs en IHM (RJC-IHM'2000, Berder, AFIHM)
- Co-responsable de l'atelier « TALN et HANDICAP » organisé dans le cadre de TALN 2001 (Tours, ATALA)
- Co-responsable de l'atelier « Méthodes hybrides pour le TALN/TALP » (TALN'1999, Cargèse, ATALA)

- Membre du comité d'organisation du congrès EURALEX'2004
- Membre du comité d'organisation du congrès TALN'2003 (Batz-sur-Mer, ATALA)
- Membre du comité d'organisation des Journées de Linguistique de Corpus (LINGCORP'01, 02 et 03, Lorient)
- Membre du comité d'organisation des Colloque *"Les relations interindividuelles médiatisées par les réseaux informatiques"* (GRESICO, Vannes, 1998)
- Membre du comité d'organisation des 2° Journées Jeunes Chercheurs en Sciences Cognitives (ARC, Gien, 1995)

4. ENCADREMENTS DOCTORAUX

Doctorat Jérôme Goulian (U. de Bretagne Sud) — Ce doctorat visait l'utilisation de techniques issues du TAL robuste ainsi que l'adaptation du formalisme des grammaires de dépendances pour la mise en œuvre d'un système de compréhension de la parole (ROMUS) en situation de dialogue homme-machine finalisé. Réalisé sous la direction officielle de Franck Poirier (VALORIA, UBS) il a été soutenu le 13 décembre 2002. J'en ai assuré l'encadrement scientifique.

Doctorat Jeanne Villaneau (U. de Bretagne Sud) — Ce doctorat étudie l'application d'approches logiques (grammaires catégorielles, λcalcul) pour la compréhension de la parole en situation de communication homme-machine. Il a donné lieu à la réalisation d'un système de compréhension (LOGUS) opérationnel implanté à l'aide du langage de programmation λProlog. L'encadrement scientifique de ce doctorat est réalisé par O. Ridoux (IRISA, Rennes) et moi-même. Il donnera lieu à une soutenance programmée pour novembre 2003.

Doctorat Igor Schadle (U. de Bretagne Sud) — L'objectif de ce doctorat est la mise en œuvre d'un système de prédiction linguistique (Sybille) destiné à l'aide à la communication pour des personnes fortement handicapées (infirmes moteurs cérébraux, tétraplégiques ayant perdu l'usage de la parole). La prédiction repose sur un modèle de langage structurel probabiliste intégrant une analyse en constituants minimaux (*chunks*). L'évaluation du système est réalisée dans le cadre d'une collaboration avec le centre de rééducation et réadaptation fonctionnelle de Kerpape. Ce doctorat se réalise sous la direction officielle de Franck Poirier (VALORIA). J'en assure l'encadrement scientifique. Il donnera lieu à une soutenance programmée pour la fin 2003.

Doctorat Mohamed Ahafhaf (U. Stendhal) — Ce doctorat concerne l'évaluation des systèmes de dialogue oral homme-machine. Il étudie l'extension aux niveaux d'interprétation contextuelle et de gestion du dialogue du paradigme d'évaluation DCR que j'avais proposé, pour la compréhension de parole, avec Jean Caelen. Ce doctorat se réalise sous la direction de Jean Caelen (CLIPS-IMAG, Grenoble). Ce co-encadrement n'a pas fait l'objet d'une déclaration auprès du conseil scientifique de l'Université Stendhal, mon intervention étant surtout présente au titre de membre extérieur du laboratoire CLIPS-IMAG et d'initiateur du paradigme DCR.

Doctorat Véronique Bralé (U. de Bretagne Sud) — L'objectif de ce doctorat est l'étude et l'intégration de connaissances permettant d'intégrer des modes d'expressivités en synthèse de parole. Le cadre applicatif retenu concerne la synthèse vocal pour les systèmes de dialogue homme-machine finalisés. Ce doctorat est réalisé pour majeure partie au sein de l'équipe synthèse de France Télécom R&D (Thierry Moudenc et Valérie Maffiolo). J'en assure l'encadrement scientifique du côté académique en compagnie de Ioannis Kanellos, directeur officiel de thèse.

Annexes C : sélection d'articles

Antoine J-Y., Goulian J. (2001) Etude des phénomènes d'extraction en français parlé sur deux corpus de dialogue oral finalisé. Application à la communication orale homme - machine. *Traitement Automatique des Langues*, *TAL*, 42(2), 413-440.

Antoine J-Y., Caelen J. (1999) Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR (Demande, Contrôle, Résultat), *Langues*, 2(2). 130-139.

Schadle I., Antoine J-Y., Le Pévédic B., Poirier F., (2002) SybiLettre, prédiction de lettres pour la communication augmentée, *Revue d'Interaction Homme-Machine, RIHM*, Vol 3, n°2, ISSN 1289-2963, Europia, Paris, France, 2002.

Villaneau J., Antoine J-Y., Ridoux O., (2001)., *Combining syntax and pragmatic knowledge for the understanding of spontaneous spoken sentences* , actes *Logical Aspects of Computational Linguistics*, *LACL'2001*, Le Croisic, France, publié in LNAI 2099, Springer, 279-295.

Goulian J, Antoine J.-Y., Poirier F. (2003) How NLP techniques can improve speech understanding : ROMUS, a robust chunk based message understanding system using link grammars. *8th European Conference on Speech Communication and Technology*, *Eurospeech'2003*.