ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

HABILITATION A DIRIGER DES RECHERCHES

SPÉCIALITÉ : Informatique

Laboratoire Informatique d'Avignon(EA 931)

Modèles numériques pour la compréhension automatique de la parole

par

Frédéric BECHET

Soutenue publiquement le 30 novembre 2007 devant un jury composé de :

M.	Jerome Bellegarda	Apple, Speech & Language Technologies, Cupertino	Rapporteur
M.	Renato De Mori	Professeur à l'Université d'Avignon	Directeur
M.	François Denis	Professeur à l'Université de Provence, Marseille	Rapporteur
M.	Marc El-Bèze	Professeur à l'Université d'Avignon	Examinateur
M.	Jean-Luc Gauvain	Directeur de Recherche, LIMSI-CNRS, Paris	Président
M.	Jean-Paul Haton	Professeur à l'Université Henri Poincaré, Nancy	Rapporteur
M.	Henri Méloni	Professeur à l'Université d'Avignon	Examinateur

---

Remerciements

Je tiens à remercier sincèrement tous les membres du jury d'avoir accepté d'évaluer ce travail résumant mes recherches conduites durant ces dix dernières années.
Je remercie particulièrement M. Jerome Bellegarda, du laboratoire Speech & Language Technologies d'Apple, M. François Denis, Professeur à l'Université de Provence et M. Jean-Paul Haton, Professeur à l'Université de Nancy, de m'avoir fait l'honneur d'être rapporteurs de mon mémoire d'habilitation.
Je remercie aussi M. Jean-Luc Gauvain, Directeur de Recherche au CNRS, pour avoir accepté de présider ce jury.

M. Renato de Mori, professeur à l'Université d'Avignon, a accepté de diriger cette habilitation, l'essentiel des résultats présentés ici est issu de travaux communs, fruits d'une active collaboration pour laquelle je lui suis tout particulièrement reconnaissant.
Si ce mémoire doit beaucoup à Renato de Mori, ma carrière de chercheur a débuté grâce à Henri Méloni, professeur à l'Université d'Avignon, et c'est dans le contexte si particulier du LIA, laboratoire qu'il a créé et amené à sa forme actuelle, que cette carrière s'est dessinée ; sa présence dans ce jury est pour moi un plaisir et un honneur.
Je tiens également à remercier Marc El-Bèze, professeur à l'Université d'Avignon, d'avoir accepté de faire partie de ce jury. Nos collaborations à l'occasion de campagnes d'évaluation ont toujours donné lieu à des discussions stimulantes et passionnées.

Je tiens également à remercier les chercheurs des laboratoires AT&T Shannon Labs et France Télécom R&D : les travaux présentés dans ce document sont pratiquement tous issus de collaborations effectuées lors de séjours ou de projets communs. Ce fut, à chaque fois, un véritable « travail d'équipe » dont je leur suis infiniment reconnaissant.

Enfin je voudrais associer à ce document, outre évidemment l'ensemble de la « famille » du LIA, les docteurs (ou futurs docteurs) dont j'ai eu le plaisir de co-encadrer les travaux de thèses et qui se retrouveront très largement dans ce document : Yannick Estève, David Janiszek, Christian Raymond, Nathalie Camelin, Christophe Servan et Frédéric Duvert.

Résumé

Cette habilitation présente mes travaux effectués au Laboratoire Informatique d'Avignon et au Shannon Labs d'AT&T dans le domaine de la modélisation du langage pour la compréhension automatique de la parole. En traitant le langage parlé, mes travaux sont à l'intersection du Traitement Automatique de la Langue Naturelle (TALN) d'une part, et d'autre part du Traitement Automatique de la Parole (TAP). Les modèles utilisés sont pour la plupart des modèles d'Apprentissage Automatique (AA) sur corpus, utilisés le plus souvent en complément de modèles à base de connaissances explicites. Une des caractéristiques principales de ces travaux est de conjuguer des problématiques de recherche académique à des cadres applicatifs définis à l'occasion de contrats de recherche avec des partenaires de l'industrie des Télécoms. Ces cadres applicatifs sont l'occasion de travailler sur des corpus *réalistes* de grande taille, permettant de tester et valider les modèles proposés.

Cette HDR tente tout d'abord de répondre à quatre questions préalables à l'appréhension du domaine de la compréhension automatique de la parole :

- Comment représenter formellement le sens d'un message ?
- Quels modèles numériques et quels algorithmes peut-on utiliser pour développer un système de compréhension ?
- Sur quels corpus et avec quelles observations peut-on entraîner un modèle numérique de compréhension ?
- Comment peut-on évaluer les performances d'un système de compréhension automatique de la parole ?

Ces questions sont aussi l'occasion d'effectuer un tour d'horizon des recherches dans le domaine.

Une fois la problématique définie, je présente une méthode de décodage conceptuel permettant de lier les processus de reconnaissance automatique de la parole et d'analyse linguistique en projetant une représentation en mots d'un message oral vers une représentation en *concepts*. Ces concepts sont les unités de base du processus de compréhension, ils sont ensuite assemblés lors d'une phase de *composition sémantique* pour produire une interprétation formelle. Différentes stratégies d'utilisation des modèles proposés sont alors présentées et évaluées.

Enfin cette HDR se conclut par l'énoncé de plusieurs principes caractérisant le domaine de la compréhension automatique de la parole et pouvant servir de recommandation au développement de modèles s'attaquant à cette tâche.

Table des matières

Chapitre 1

Introduction

Sommaire

1.1 Avant-propos

Ce document présente mes travaux réalisés dans le cadre du Traitement Automatique de la Parole (TAP), et plus particulièrement ceux concernant la compréhension automatique de messages oraux. Ces travaux ont débuté il y a une dizaine d'années, à travers le co-encadrement avec le Professeur Renato de Mori, d'une thèse portant sur la modélisation du langage dans le cadre du dialogue oral. Cette thèse, effectuée par Yannick Estève aujourd'hui enseignant-chercheur à l'Université du Mans, s'est faite à l'occasion d'un premier partenariat avec le laboratoire de recherche de France Télécom à Lannion, le *CNET*, laboratoire incontournable dans l'histoire du TAP en France (et qui depuis change régulièrement de nom ...). Ce contrat a été la première occasion d'être confronté à des données *réalistes* de dialogue homme-machine, d'envisager des tâches dépassant le paradigme de la dictée vocale qui reste le modèle dominant en Reconnaissance Automatique de la Parole (RAP), et enfin d'appréhender le problème de la communication homme-machine sur des traces *écologiques* et non plus sur des exemples artificiels de laboratoire. Ce travail sur les données, les corpus contenant de la parole complètement spontanée collectée grâce à des applications mises en service,

a continué lors des collaborations successives avec France Télécom : depuis les corpus contenant quelques centaines de messages collectés lors d'expérimentations en conditions réalistes pour la première convention, jusqu'aux corpus de taille potentiellement infinie collectés de manière journalière auprès d'applications mises en service sur une large échelle.

Ce sont mes premiers travaux effectués sur la modélisation du langage dans le cadre du dialogue oral qui m'ont ensuite permis d'effectuer un séjour d'un an au laboratoire américain *AT&T Shannon Labs* à Florham Park dans le New Jersey, de septembre 2001 à septembre 2002. Durant ce séjour j'étais intégré au sein de l'équipe de recherche responsable du projet *How May I Help You ?* dirigé par Allen Gorin. Cela a été l'occasion de travailler sur l'application de dialogue sans doute la plus largement déployée à ce jour. En effet elle couvrait l'ensemble du territoire américain, traitant plusieurs millions d'appels par mois. Même si l'application de dialogue était assez limitée, pour des raisons de robustesse, le corpus collecté à cette occasion contient une très grande variété de phénomènes due au nombre et à l'extrême diversité des locuteurs utilisant le service. Cette expérience m'a permis notamment d'appréhender les limites des techniques actuelles de RAP, confrontées à des difficultés absentes des corpus de laboratoires ainsi que des corpus *statiques* d'évaluation sur lesquels les systèmes sont optimisés année après année.

L'ensemble des travaux présentés dans ce document a donc été réalisé lors de partenariats avec des laboratoires de recherche issus de l'industrie des télécoms. Ces partenariats ont été l'occasion unique d'étudier des données inaccessibles par ailleurs, et d'être confronté à des cadres applicatifs concrets, qui sont bien souvent les seules justifications des modèles numériques mis en œuvre comme je le développerai par la suite. C'est pourquoi ce document mettra l'accent sur l'importance et la description de ces corpus, le TAP fondé sur des modélisations empiriques étant par essence un domaine de recherche *expérimental*.

Un autre aspect important de ces travaux est l'évaluation. La nécessité d'évaluer les modèles proposés est devenu un paradigme incontournable en TAP depuis les années 1980, et le souci de comparer les méthodes a toujours été constant dans ces travaux, illustré par ma participation à de nombreuses campagnes d'évaluation françaises (Technolangue/Evalda avec les campagnes ESTER, EVASY et MEDIA de 2003 à 2005 ; la campagne DEFT en 2005 et 2007) et internationales (Pascal Challenge en 2005, Document Understanding Conference, DUC, en 2006 et 2007). Enfin ces travaux sont le fruit de collaborations fructueuses avec des partenaires de recherche de nombreux laboratoires français et étrangers, collaborations ayant souvent pour cadre les contrats de recherche sur lesquels j'ai eu l'occasion de travailler, contrats industriels comme présenté précédemment mais aussi contrats académiques lors de la participation à des projets français (Technolangue de 2003 à 2005, projet ANR EPAC depuis janvier 2007) et européens (projet STREP SMADA de 2000 à 2003, projet STREP LUNA depuis septembre 2006).

1.2 Cadre de l'étude

Cette étude se situe dans le cadre des modèles numériques de traitement du langage. Le but ici n'est pas de développer un modèle explicatif d'un phénomène linguistique mais de reproduire une compétence linguistique particulière. Cette compétence étant représentée par une séquence d'observations, elle est donc *observable*. Ces deux points sont fondamentaux pour la suite :

1. Nos modèles n'ont pas d'ambition explicative ou descriptive, ils modélisent un phénomène *naturel*, similaire en ce sens, par exemple, aux modèles météorologiques : des observations et une *interprétation* de ces observations sont disponibles jusqu'à l'instant t, le but est de prédire l'interprétation des observations obtenues à l'instant $t + 1$.
2. Pour ce type de modèles les compétences modélisées doivent être *observables*, ce point va nous permettre de définir de quel type de *compréhension* il sera question dans ce document.

Nous allons détailler les conséquences de ces deux choix.

1.2.1 Modèles numériques

Puisque les modèles développés n'ont pas pour but de valider une théorie linguistique ou d'expliciter un phénomène, toute liberté est donnée dans leur conception. Les paramètres utilisés, leurs modélisations, les stratégies de décision n'ont pas à être motivés par autre chose que leur performance par rapport à la compétence linguistique reproduite. Cette liberté vis-à-vis des théories linguistiques a ainsi son pendant : leur seule justification est leur performance, elle doit donc être évaluable de manière fiable. Cette évaluation nécessite un protocole expérimental et une méthodologie d'évaluation, nous verrons qu'il s'agit là d'un des problèmes principaux de ce type de méthodes.

1.2.2 Compétence linguistique observable

Le paradigme des modèles numériques considérés est la modélisation d'un phénomène naturel afin de pouvoir effectuer des prédictions sur l'interprétation d'événements futurs. Ce paradigme nécessite de disposer d'observations, ce sont les *traces* du phénomène à modéliser. A partir de ces traces le phénomène peut être caractérisé (ou *interprété*), et en disposant des observations et de leur caractérisation (ou *interprétation*) jusqu'au temps t, le modèle doit pouvoir prédire le comportement du phénomène dans le futur ainsi que ses caractéristiques. Par exemple, toujours en prenant l'analogie avec les modèles numériques météorologiques, si le phénomène à observer est l'apparition de tornades, les observations peuvent être les paramètres de pression atmosphérique, la température et la vitesse des vents. La caractérisation du phénomène à partir des observations, son *interprétation*, est la présence ou pas de tornades, sa violence, etc.

Si l'on garde le même paradigme dans le cadre de cette étude, c'est à dire la reproduction par une machine d'une compétence linguistique particulière, il faut nécessairement que cette compétence laisse des *traces* du phénomène sous-jacent, ces traces constituant les interprétations des observations prises en entrée par les modèles numériques. Ainsi, cette étude portant sur la notion de *compréhension de la parole*, nous allons nous intéresser à des aspects du processus de compréhension pouvant être observés directement. Cette observation va se faire sur des *objets linguistiques* représentant des messages oraux.

Ces *traces* de compréhension sont de deux sortes : tout d'abord des traces *directes* où l'objet linguistique contient déjà des indications liées à la compréhension (par exemple les actions d'un système de dialogue en langage naturel) et les traces *indirectes* où on demande à un expert possédant la compétence à reproduire, appelé *annotateur*, d'indiquer par un code ces traces relatives à la compréhension.

1.3 Vue d'ensemble des thèmes abordés

L'objectif de cette section est de décrire le contexte scientifique dans lequel s'insère cette habilitation, située à l'intersection de trois thématiques : l'Apprentissage Automatique (AA), le Traitement Automatique des Langues Naturelles (TALN) et la Reconnaissance Automatique de la Parole (RAP). Les prochains paragraphes présentent brièvement les liens entre ces thématiques et introduisent le thème de la compréhension automatique de la parole.

1.3.1 Traitement automatique des langues et apprentissage automatique

Le Traitement Automatique des Langues Naturelles (TALN) a longtemps évolué en parallèle avec la linguistique formelle, en particulier depuis les travaux fondateurs de Noam Chomsky dans les années cinquante (Chomsky, 1957). La division du travail entre les deux disciplines était simple : la linguistique formelle fournissait les outils formels pour la description des langues (que l'on désigne habituellement par le terme de formalismes linguistiques) tandis que le TALN s'attelait à décrire et implémenter des processus reposant sur ces formalismes afin de réaliser certaines tâches, dont l'exemple le plus illustre est la traduction automatique. Certaines faiblesses de ce mode de fonctionnement sont apparues au fil du temps et se sont cristallisées autour de trois points : la constitution de ressources, le traitement de l'ambiguïté et la robustesse.

- La constitution de ressources désigne la tâche visant à décrire une langue (en particulier sa syntaxe et son lexique) dans le cadre d'un formalisme linguistique. Il s'agit d'une tâche extrêmement coûteuse (elle mobilise de nombreux linguistes pendant plusieurs années), rarement menée à son terme du fait de l'instabilité des formalismes (ils évoluent en même temps que le travail de développement).
- L'ambiguïté est un phénomène intimement lié à la langue qui permet d'associer plusieurs *structures* à un même énoncé. Si un être humain peut généralement

désambiguïser sans peine la plupart des énoncés, il n'en est pas de même d'un ordinateur, qui ne possède qu'une petite partie des connaissances mobilisées par un locuteur pour mener à bien cette tâche. En effet, la langue naturelle étant par essence implicite et ambiguë, ce n'est que par la masse des connaissances partagées entre les locuteurs que cette ambiguïté peut être levée. L'effet est qu'à l'issue d'un traitement informatique, on aboutit généralement à plusieurs interprétations et que l'on ne dispose pas des moyens permettant de choisir la bonne.

– La robustesse désigne la capacité d'un système à traiter des entrées bruitées (dans le cas du TALN, le bruit est principalement constitué des fautes d'orthographe et de syntaxe, de phrases incomplètes ou de la présence de marques extra-textuelles). Les systèmes de TALN réagissent mal à ce genre de phénomènes. En effet, ils reposent souvent sur des modèles fins de la langue (en partie pour essayer de limiter le phénomène de l'ambiguïté) et échouent sur des entrées bruitées.

Parallèlement à cette évolution du TALN, une autre voie s'est développée, fondée sur la notion de *méthodes empiriques*, visant à résoudre certaines tâches linguistiques sans recourir à la linguistique formelle. Cette approche se distingue fondamentalement de la première dans son rejet des formalismes linguistiques complexes. Elle repose sur des modèles linguistiques simplistes, appris automatiquement à partir de données. Cette notion d'*apprentissage* à partir des données a permis à la communauté TALN d'établir des liens clairs avec le domaine de l'Apprentissage Automatique qui s'intéresse à la conception d'algorithmes et de techniques permettant à un ordinateur d'*apprendre* à résoudre certains problèmes grâce à des exemples qui lui sont fournis en entrée. Aujourd'hui les liens entre les deux disciplines sont solidement établis, et constituent un domaine de recherche extrêmement actif.

Ces modèles empiriques offrent des solutions aux trois problèmes mentionnés ci-dessus : leur constitution est automatique sous réserve de la disponibilité de corpus d'apprentissage (ce qui est bien souvent un problème crucial, comme nous le verrons dans le chapitre suivant) et ils fournissent un score aux différentes interprétations d'un énoncé, permettant ainsi de résoudre l'ambiguïté en choisissant l'interprétation ou les quelques interprétations les mieux notées. Ces scores permettent aussi d'intégrer facilement les différents processus linguistiques impliqués dans la réalisation d'une tâche en définissant un *espace de recherche* contenant toutes les interprétations obtenues à chaque niveau. La décision finale peut être vue comme une recherche de meilleur chemin dans cet espace.

De plus, ces modèles empiriques reposent sur des modèles très peu contraints de la langue (en particulier de la syntaxe) ce qui leur permet de mieux réagir face à des phrases agrammaticales ou à la présence de mots mal orthographiés. Les techniques développées dans ce cadre ont petit à petit été appliquées aux différentes tâches du TAL : étiquetage morpho-syntaxique (Merialdo, 1994), analyse syntaxique (Charniak, 1997), traduction automatique (Brown et al., 1990) ...

Un des paradigmes utilisé par les méthodes empiriques pour modéliser la communication homme-machine est fondé sur les travaux de Shannon concernant les principes de la théorie de l'information (Shannon, 1948). Ce paradigme est à la base des modèles

de RAP comme nous le verrons dans le paragraphe suivant.

1.3.2 Reconnaissance Automatique de la Parole

La reconnaissance automatique de la parole (RAP), dans un sens très général, est la tâche consistant à traiter à l'aide d'un ordinateur un enregistrement vocal. Cette tâche se décline en sous-tâches plus précises, ainsi les avancées récentes des technologies de reconnaissance vocale ont permis le développement de systèmes opérationnels illustrant les deux grandes fonctions de la langue naturelle : outil de représentation des connaissances d'une part ; outil de communication d'autre part. Dans la première famille d'applications on trouve les systèmes de dictée vocale, d'indexation et de structuration de documents audio. Les systèmes de dialogue oral entre une personne et une machine, représentés essentiellement par les serveurs vocaux téléphoniques, constituent la deuxième famille d'applications.

Quelles que soient les applications ou les langues visées, la totalité des systèmes de RAP développés actuellement implémentent une approche probabiliste, fondée sur la théorie de l'information (Shannon, 1948). Dans cette approche, la parole est vue comme un canal bruité entre une source émettant une séquence d'*événements linguistiques* (mots, sens, actes communicatifs,...) et un récepteur devant débruiter le signal reçu pour obtenir la séquence d'événements émise. Cette modélisation probabiliste est exprimée par la probabilité $P(W|A)$ représentant la probabilité de la séquence de mots W émise étant donnée la séquence d'observations acoustiques A représentant le signal vocal. Deux types de modèles sont utilisés pour estimer cette probabilité : d'une part les modèles acoustiques permettant de donner une probabilité à une suite de mots d'avoir été prononcée étant donné un signal vocal ; d'autre part les modèles linguistiques, appelés aussi modèles de langage ou modèle de langue, estimant la probabilité qu'une séquence de mots appartienne au langage modélisé.

Les performances des systèmes de RAP sont généralement évaluées par le critère du *taux d'erreur par mot* (ou *taux d'erreur mot*) mesurant la distorsion entre la séquence de mots émise par la source et celle décodée par le récepteur. Un taux de 0% correspond à une transmission parfaite entre la source et le récepteur. Dans le cas de la parole lue la séquence de mots émises de référence est constituée directement par le texte lu. Dans le cas de parole non lue, une référence est obtenue en faisant écouter le message vocal à un auditeur humain et en lui demandant de transcrire la séquence de mots perçue.

Les résultats annoncés comme étant représentatifs de l'*état de l'art*, dans les campagnes d'évaluation telles que celles de NIST HUB5 aux Etats-Unis (Pallet, 1997; Fiscus et al., 2000) ou Technolangue/Evalda/ESTER (Gallliano et al., 2005) en France, varient selon le type d'application visée : environ 5% d'erreurs pour la parole lue, 10% pour la transcription de données radiodiffusées (journaux radiophoniques) et entre 20 et 30% d'erreurs pour la parole spontanée téléphonique (dialogue homme-machine ou conversation téléphonique).

Ces taux d'erreurs sont des taux moyens cachant une grande disparité de performances selon les messages traités, les facteurs de difficultés de traitement d'un message

vocal étant principalement :

- Les mauvaises conditions acoustiques telles que les bruits ambiants, les dégradations dues au canal de transmission (téléphone portable par exemple), ou encore les signaux ajoutés au signal de parole à reconnaître (musique ou locuteurs multiples).
- Les voix non représentées dans les bases d'apprentissage des modèles acoustiques, par exemple les accents régionaux ou de locuteurs non natifs ou encore les voix dégradées à la suite d'une pathologie ou encore lors de l'expression d'une émotion particulière (stress, colère, etc.).
- Le manque de couverture des modèles linguistiques : absence d'un mot à reconnaître dans le lexique du système de reconnaissance, ou encore absence d'une structure syntaxique dans le modèle grammatical de ce même système.

Les applications fondées sur ces technologies commencent à être largement répandues dans des domaines comme la dictée vocale, les serveurs téléphoniques tels que les applications de renseignement téléphonique (Boves et al., 2000) ou de service clientèle (Damnati et al., 2007a; Gorin et al., 1997), ou encore les applications d'indexation de documents audio pour des tâches de recherche documentaire comme dans le projet RNTL AUDIOSURF.

Ces systèmes sont opérationnels lorsque les conditions d'utilisation sont proches de celles définies lors de la phase d'apprentissage des modèles de reconnaissance. Ceci constitue la première limitation des technologies actuelles due à un verrou scientifique important : comment prendre en compte un nouveau contexte d'utilisation non vu lors de la phase d'apprentissage des modèles ?

Une deuxième limitation concerne l'interface entre un système de RAP et les modules de traitement des transcriptions automatiques produites. En effet, la retranscription n'est souvent pas une fin en soi[1] : les sorties d'un module de RAP sont en général fournies en entrée à un système de compréhension dont le but consiste à produire une représentation du sens de l'énoncé, ou d'une partie de ce dernier. Or actuellement, cette articulation est loin d'être optimale. Il s'agit là du point de contact entre les deux disciplines que sont la reconnaissance de la parole et le TALN. Cette interface constitue une des problématiques de cette habilitation.

1.3.3 Compréhension de la parole spontanée

La tâche de compréhension de la parole spontanée repose sur les deux domaines du TALN et de la RAP présentés ci-dessus. Elle consiste à traiter un enregistrement vocal afin de construire une représentation partielle de son sens, utile pour une application donnée. Une première manière d'envisager de résoudre ce problème consiste à prendre en entrée d'un module de compréhension, la retranscription fournie par un système de RAP. Cette approche séquentielle du traitement automatique de messages oraux, dans laquelle les transcriptions automatiques constituent la seule interface entre les modèles

[1] A l'exception de la tâche de dictée vocale.

de reconnaissance de la parole d'une part et les modules d'analyse linguistique d'autre part est inadéquate pour les raisons suivantes :

1. Contrairement aux documents écrits, la retranscription d'un message oral ne constitue qu'une partie de l'information véhiculée par ce message. En effet, si c'est une évidence d'affirmer qu'un document écrit est destiné à être lu, il n'en est évidemment pas de même pour un message oral où les informations prosodiques (débit, intonation, énergie) sont souvent indispensables (par exemple pour reconnaître une question exprimée sous une forme affirmative). Les informations absentes de la retranscription peuvent également être pertinentes pour caractériser un message. Parmi ces informations on peut trouver les silences, les disfluences (mots tronqués, reprises, hésitations, corrections, ...) où même l'environnement sonore tel que les *jingles* dans le cadre du traitement de données audiovisuelles.
2. La production d'énoncés oraux ne répond pas aux mêmes normes que les énoncés écrits (phrases non achevées, structures agrammaticales, répétitions, absence de marques de structure formelle telles que la ponctuation). Il est par conséquent illusoire d'analyser des transcriptions automatiques à l'aide d'outils conçus pour traiter de la langue écrite.
3. Les performances des systèmes de RAP sur de la parole spontanée restent assez faibles. A titre d'illustration on peut citer les résultats récents obtenus sur des tâches de transcription automatique d'enregistrement audio de réunions de travail (Hain et al., 2005), le traitement de très vastes corpus de témoignages audio comme dans le projet MALACH (Ramabhadran et al., 2003), ou encore l'analyse de messages collectés dans le cadre de sondages d'opinions présentés dans le chapitre 3 (Camelin et al., 2006). Dans ces trois exemples les taux d'erreur mot des meilleurs systèmes dépassent souvent les 50% d'erreurs. Les raisons de ce taux d'erreurs important sont profondes et on ne peut espérer des améliorations substantielles dans un proche avenir. Une manière de pallier le problème consiste à produire non plus la seule retranscription, jugée la plus probable par le système de RAP, mais un ensemble de retranscriptions qui seront fournies au module de compréhension.
4. La tâche de transcription littérale peut être considérée comme artificielle dans le traitement de messages spontanés où le contexte d'émission et les informations acoustiques (prosodiques, événements non linguistiques) sont indispensables à la compréhension du message en complément de l'information linguistique. Ainsi c'est bien la tâche de *compréhension* qui est pertinente et non celle de transcription.

Cette notion de compréhension doit être définie en fonction des contextes d'élocution et d'utilisation des messages vocaux. Par exemple, dans un cadre de dialogue homme-machine, la compréhension d'un message nécessite sa traduction dans une représentation formelle utilisée par le gestionnaire de dialogue. Pour une tâche d'indexation de corpus, cette notion de compréhension peut être représentée par la détection du thème, des *concepts* tels que les entités-nommées, les actes dialogiques ou encore les rôles sémantiques présents dans le message.

Ainsi on peut définir la *compréhension* comme l'ensemble des analyses visant à caractériser, étiqueter, structurer et finalement représenter formellement l'information contenue dans un message en fonction des contextes d'élocution et d'utilisation de ceux-ci.

1.4 Plan du document

Le document est structuré comme suit :

- Le chapitre 2 introduit la problématique de la compréhension automatique de la parole en posant quatre questions fondamentales :
  1. Comment représenter formellement le sens d'un message ?
  2. Quels modèles numériques et quels algorithmes peut-on utiliser pour développer un système de compréhension ?
  3. Sur quels corpus et avec quelles observations peut-on entraîner un modèle numérique de compréhension ?
  4. Comment peut-on évaluer les performances d'un système de compréhension automatique de la parole ?
- Les corpus utilisés dans cette étude sont tous présentés au chapitre 3.
- Dans le chapitre 4 nous présentons une méthode de *décodage conceptuel* permettant de projeter une représentation en mots d'un message vers une représentation en *concepts*, ces concepts représentant les unités de base du processus de compréhension.
- Une fois les unités de sens extraites il convient de les *composer* afin de donner une interprétation globale à un message oral, cette étape de *composition sémantique* est le sujet du chapitre 5.
- Différentes stratégies d'utilisation des modèles proposés sont présentées et évaluées dans le chapitre 6.
- Enfin le chapitre 7 conclut ce document en énonçant un certain nombre de principes ayant pour but de définir le domaine de la compréhension automatique de la parole et servir de recommandation au développement de systèmes s'attachant à ce domaine.

Chapitre 2

Quatre questions préalables à la compréhension automatique de la parole

Sommaire

2.1 Comment représenter le sens d'un message oral ?

2.1.1 Représentation du sens dans un cadre applicatif

Le problème de la représentation du *sens* d'un message oral est un point d'étude qui peut être envisagé selon de nombreuses perspectives : philosophique, linguistique, cognitive, mathématique ou computationelle. Dans le contexte de l'application de méthodes numériques à la tâche de compréhension automatique de la parole, la principale justification des méthodes présentées dans ce document est leur performance dans un cadre applicatif, leur capacité à *reproduire* une compétence linguistique par rapport à une tâche donnée.

C'est dans cet esprit que le problème de la représentation du sens va être abordé : plutôt que de choisir une théorie sémantique donnée pour envisager ce problème, nous l'aborderons sous l'angle pragmatique de son utilisation dans un cadre applicatif. Le *sens* d'un message est donc ici considéré comme l'ensemble des informations nécessaires à un système pour réaliser une tâche donnée sur le message pris en entrée. Ce sens n'a donc pas de portée générale hors du cadre de l'application.

Nous nous situons ici dans le cadre de la *sémantique procédurale* résumée par Woods (Woods, 1981, 1983) dans la citation suivante reprise par (Schirra et Joerg, 1993) :

« ... that the meaning of a symbol resides in an abstract procedure, not necessarily executable, linking the symbolic expression to the physical world through the (computational / inferential) operations of a physical interpreter operating on a combination of internal representations and sensory/motor connections to the world. » [1]

Cette définition permet d'envisager le sens d'une manière procédurale, le monde physique est ici l'état du système traitant le message, l'interpréteur est le module de compréhension qui va faire l'interface entre d'une part les observations acoustiques reçues en entrée et représentant le message vocal et d'autre part les fonctionalités du système intégrant ce module de compréhension.

Par exemple, dans une tâche de dialogue oral homme-machine, le sens de l'intervention d'un utilisateur peut être représenté par l'action que doit effectuer la machine en réponse à celle-ci. Ainsi l'interprétation du message : « *je veux connaître le solde de mon compte* » peut-être représentée par la requête à la base de données permettant d'extraire l'information recherchée. C'est cette représentation du sens qui avait été choisie dans la campagne d'évaluation des systèmes de compréhension ATIS (Pallett et al., 1992), dans le domaine des systèmes de dialogue pour la réservation aérienne, où le sens de chaque message d'utilisateur était représenté par la requête SQL lui correspondant.

Un autre exemple est celui des systèmes de *routage d'appel* (ou *call routing*) tels que le système d'AT&T *How May I Help You ?* (Gorin et al., 1997). Dans de tels systèmes le sens

[1] ... le sens d'un symbole réside dans une procédure abstraite, qui n'est pas nécessairement exécutable, liant l'expression symbolique au monde physique à travers des opérations de calcul et d'inférence. Ces opérations sont effectuées par un interpréteur agissant sur une combinaison de représentations internes et de connections sensori-motrices avec le monde physique.

d'un message est la destination vers laquelle le message doit être dirigé : sous-menu de dialogue, opérateur humain, message enregistré, etc.

Cette représentation *ad hoc* du sens pour un cadre applicatif précis a l'avantage d'être directement liée à l'utilisation de ce sens pour réaliser la compétence linguistique qu'est censée simuler la machine. Ainsi une amélioration des performances de compréhension telle qu'elle est définie ici aura un impact direct sur les performances du système global.

Cependant ce lien très fort entre « *sens* » et « *contexte applicatif* » comporte deux inconvénients majeurs : il ne permet pas de faire de l'évaluation détaillée des capacités d'un système et il nécessite la mise en place d'un système complet avant de pouvoir faire la moindre évaluation. En étant trop proche d'un système précis, les représentations *ad hoc* du sens manquent de généricité et le développement d'une nouvelle application nécessite de redéfinir complètement un nouveau système d'annotation. De plus les systèmes complets susceptibles d'être mis en service actuellement sont généralement de faible complexité (routage d'appel et remplissage de formulaires par exemple), ce qui entraîne une représentation du sens elle aussi de faible complexité.

Pour répondre à ces deux inconvénients une autre approche consiste à utiliser une représentation du sens liée à une théorie sémantique, la sous-section suivante présente les approches de ce type les plus utilisés en traitement automatique de la langue.

2.1.2 Représentation du sens à partir d'un modèle linguistique

La représentation du sens par une machine doit être exprimée par un langage formel qui a sa propre syntaxe et sémantique. Ce langage doit être cohérent avec une théorie sémantique donnée (représentation en extension ou en intension, traits sémantiques, relations, raisonnement, composition de constituants sémantiques, liaison entre sens et formes de surface, *etc*.). L'extraction d'un sens formel d'un message est généralement envisagée sous l'angle de la *sémantique compositionnelle* qui vise à *composer* des objets sémantiques de base, sous la forme de règles logiques, pour extraire un sens global à un énoncé tout en se détachant au maximum de ses formes de surfaces.

Du point de vue du système, les connaissances sémantiques pouvant être traitées par une application sont une *base de connaissances* que l'on peut représenter de manière pratique par un ensemble de formules logiques. Ces formules contiennent des variables qui peuvent être instanciées par des constantes et qui peuvent être typées. Un *objet sémantique* peut être construit en instanciant toutes les variables d'une formule ou en composant entre eux différents objets. Le rôle du processus d'interprétation est d'instancier de tels objets et de déduire de nouveaux faits par un processus d'inférence. La logique du premier ordre est généralement utilisée pour représenter de telles bases de connaissances.

Représentation logique

Suivant les travaux de Woods (Woods, 1975) sur la représentation sémantique, l'ensemble de formules logiques constituant la base de connaissances d'un système peut représenter un *réseau sémantique* définissant des entités et des relations entre ces entités. Un réseau sémantique est fait de nœuds correspondant à des *concepts* et de relations entre ces nœuds. Par exemple ces relations peuvent être des relations de compositions comme dans les travaux de Jackendoff (Jackendoff, 1990). Plusieurs langages ont été proposés pour représenter ces réseaux sémantiques, la plupart dérivés des travaux de Brachman et du langage KL-ONE (Brachman, 1979).

Les *concepts*, nœuds des réseaux sémantiques, représentent les «*briques*» à partir desquelles l'interprétation d'un message peut être faite. Chaque concept contient un ensemble d'attributs le définissant. Ces attributs peuvent décrire des parties du concept (par exemple *doigts* pour le concept *main*), des caractéristiques physiques (par exemple la couleur) ou bien encore définir des relations avec d'autres concepts. Dans ce cas ces attributs sont appelés des *rôles sémantiques* qui expriment une relation de sens d'un constituant vers un autre constituant. Par exemple la notion d'*agent* pour un verbe. Les concepts fondés sur des verbes représentent un composant fondamental de l'expression du sens d'un message. Les rôles sémantiques attachés à un verbe permettent de définir son sens par rapport à un contexte d'utilisation.

La représentation logique usuelle d'un verbe et de ses rôles est la structure prédicat/argument. Par exemple, la phrase «*l'éditeur publie des livres*» peut être représentée par : `publier(éditeur,livres)`. Chaque argument peut être une structure complexe contenant des rôles et des valeurs, par exemple un éditeur peut avoir un nom et une adresse, représentée aussi par des prédicats : `nom(éditeur,Dupont)`.

Cadres sémantiques

Du point de vue de l'implémentation, la représentation des concepts et relations peut être faite par un langage à base de *cadres sémantiques* (*semantic frames*) tel que celui proposé par Fillmore (Fillmore, 1985). Par exemple la représentation d'une adresse peut être faite par le cadre suivant :

```
{a0001
instance_of     address
  loc             Avignon
  area            Vaucluse
  country         France
street          1, avenue Pascal
code            84000
}
```

Ici `a0001` représente l'identifieur d'une instance de la classe `address`. Les autres champs de cette classe définissent les propriétés de l'objet. Ces structures sont représen-

tées formellement par des *grammaires de cadres* (ou *frame grammars*). Ces grammaires génèrent des cadres contenant des concepts généraux et des instances spécifiques. Chaque champ d'un cadre décrit un attribut ou un rôle du concept représenté par le cadre. Ces rôles et attributs peuvent à leur tour être des instances d'autres cadres sémantiques. Cette génération d'instances d'entités et de relations peut être définie dans le cadre de la *sémantique procédurale* telle que définie par Woods (Woods, 1981). Dans ce cas, des procédures sont attachées aux champs des cadres avec les préconditions nécessaires à leur exécution. Elles permettent de remplir les champs et d'effectuer des inférences sur les objets déjà obtenus.

Un verbe et ses rôles peuvent aussi être représentés par des cadres sémantiques comme dans l'exemple suivant :

```
{accept
is_a          verb
subject       [human..]
theme    [....]
....
Other roles [....]
}
```

Les termes entre crochets représentent des contraintes sur les types de valeurs pouvant être obtenues. Par exemple, une instance d'un cadre pour le verbe *accepter* représentant le prédicat `accepter(user,service004}` peut être :

```
{V004
instance_of    accepter
subject        user
theme          [service_004]
....
Other roles    [....]
}
```

Il n'existe pas de référence absolue contenant la liste des concepts de type verbe avec pour chacun d'eux la liste des rôles sémantiques lui étant associé. Comme précisé dans le paragraphe précédent on peut se situer plus ou moins près de la tâche à accomplir pour préciser les concepts, les attributs et les rôles utiles à la modélisation d'un problème. Un certain nombre d'études ont cependant essayé de construire des ressources *génériques* pouvant être utiles à la réalisation de n'importe quelle tâche. Parmi ces études une des plus célèbres est le *répertoire de cas* de Fillmore (Fillmore, 1968). Un nombre restreint de cas (moins d'une dizaine) sont définis (*agent*, *siège*, *objet*, *instrument*,...), chaque cas représentant un rôle pour n'importe quel concept de type verbe.

Dans la même lignée un formalisme tel que celui de *PropBank* (*Proposition Bank*) (Kingsbury et Palmer, 2003) propose de décrire chaque verbe selon une structure prédicat/argument. Ce type d'annotations peut être utilisé pour entraîner des systèmes d'apprentissage supervisé permettant d'étiqueter automatiquement du texte avec des rôles sémantiques. Cette annotation est fondée sur des arbres syntaxiques tels que ceux du

corpus *TreeBank* (Marcus et al., 1994). Les rôles utilisés pour décrire les verbes sont de deux sortes : des rôles généraux inspirés des cas de Fillmore et des rôles spécifiques à chaque verbe. Notons que l'annotation avec PropBank nécessite une analyse syntaxique profonde et se trouve ainsi peu adaptée au traitement de messages oraux spontanés dont une telle analyse n'est pas toujours possible à cause des phénomènes de disfluence et d'agrammaticalité comme il sera présenté dans le paragraphe 2.2.4.

Un autre formalisme, moins lié à la syntaxe, est celui proposé à Berkeley par le projet *FrameNet* (Baker et al., 1998). Ce projet tente de répondre à la question suivante : comment définir un ensemble de rôles qui ne soit ni trop général (et donc peu effectif), ni trop spécifique (et donc peu généralisable) ? L'approche proposée consiste à partir de l'étude de corpus en utilisant une méthodologie bien définie pour proposer et définir de nouvelles *Frames* quand l'étude d'un nouveau phénomène sur corpus ne permet pas d'utiliser les *Frames* existantes. Chaque *Frame* contient des éléments (*Frame Element*) tous décrits par des unités lexicales. Par exemple la *Frame* `Commerce_vente` va contenir les *Frame Elements* suivants : `acheteur`, `vendeur`, `prix`, `objet`. L'annotation *FrameNet* est une annotation partielle, seulement certaines parties d'une phrase peuvent être étiquetées.

Actes de dialogues

Les structures prédicatives sont une composante nécessaire (mais pas suffisante) de l'annotation sémantique, particulièrement pour les messages énoncés dans un cadre de dialogue. Par exemple, le message suivant : « à Marseille », du point de vue des cadres sémantiques, représente uniquement l'expression d'un lieu. Cependant dans un contexte de dialogue il peut avoir les significations suivantes :

1. une réponse à une question « Ou voulez-vous aller ? » ;
2. un choix dans une liste « Je vous propose un hôtel à Aix et un à Marseille » ;
3. une confirmation « Vous voulez un hôtel à Marseille ? » ;
4. une négation suivie d'une correction « Vous avez choisi un hôtel à Aix »
5. une question « Plusieurs hôtels dans plusieurs villes ont des disponibilités aux dates voulues ».

Ces différentes interprétations ne peuvent être obtenues qu'en prenant en compte le *contexte du dialogue*. Cette notion de contexte peut englober différents niveaux, par exemple H. Bunt (Bunt, 1995) distingue cinq types de contextes, certains fixés au début de l'interaction et d'autres changeant au cours de l'interaction :

1. le contexte linguistique, qui inclut à la fois la langue parlée par les interlocuteurs et aussi les tours de parole précédents ;
2. le contexte sémantique, formé par les objets relatifs à la tâche ;
3. le contexte physique, caractérisé par le lieu et le temps où se situe l'interaction ;
4. le contexte social, qui couvre le type de situation interactive et les rôles des participants dans cette situation ;

5. enfin le contexte cognitif qui comprend les attitudes, les objectifs et les croyances des intervenants.

D'après Bunt, repris dans (Jermann, 1996), certains aspects de ce contexte de dialogue vont être modifiés par des *éléments de comportement communicatif* :
« *..., a dialogue is viewed as a sequence of complex elements of communicative behaviour, intended to change the dialog context ...* » [2]

Ces éléments, appelés des *actes de dialogue*, peuvent être définis comme des *fonctions communicatives* permettant de changer le contexte étant donné le contenu sémantique et la forme de l'énoncé.

De nombreuses études ont proposé des schémas d'annotation pour ces *actes de dialogue*. Par exemple le projet Verbmobil (Alexandersson et al., 1998), le manuel d'annotation du projet HCRC (Carletta et al., 1997), les projet DAMSL (Core et Allen, 1997) et SWBD-DAMSL (Jurafsky et al., 1997) ou plus récemment le projet ICSI-MRDA (Dhillon et al., 2003).

Le projet Européen LUNA (Raymond et al., 2007), qui est le cadre de plusieurs travaux présentés dans ce document, utilise un ensemble restreints d'actes de dialogue provenant du projet DAMSL. Il est constitué de 8 actes principaux, 4 ayant une fonction d'anticipation (Forward looking function), et 4 ayant une fonction de retour en arrière (Backward looking function) :

- Forward looking function
  1. Statement
  2. Action-directive/open option
  3. Committing-speaker-future-action
  4. Info-request
- Backward looking function
  1. Answer
  2. Accept
  3. Reject
  4. Signal-understanding/Signal-non-understanding

2.1.3 Discussion

Parmi les très nombreux modèles qui ont été proposés pour représenter le sens d'un message, les représentations utilisées dans cette étude se justifient par rapport aux quatre points suivants :

[2] *un dialogue peut être vu comme une séquence d'éléments complexes de comportement de communication, ayant pour but de changer le contexte du dialogue*

Représentation « à plat » *vs.* représentation structurée

Le premier point concerne la complexité de la représentation nécessaire à l'exécution d'une tâche donnée. En effet la plupart des représentations utilisés dans ce travail (chapitre 4) sont des représentations « à plat » dans lesquelles la représentation du sens d'un message est constitué de la séquence d'entités de base le définissant, sans composition entre elles. Cette représentation simplifiée est généralement suffisante pour un grand nombre d'applications de type routage, classification de message ou interrogation de bases de données.

Liaison sémantique/syntaxe

Les trois derniers points sont spécifiques à la problématique du traitement de messages oraux contenant de la parole spontanée. Comme nous le verrons dans le paragraphe 2.2.4, celle-ci se caractérise par des disfluences et une forte agrammaticalité. De plus la contrainte de travailler sur des transcriptions automatiques produites par des reconnaisseurs de parole pouvant faire de nombreuses erreurs limite la profondeur de l'analyse pouvant être effectuée de manière robuste. Pour ces raisons nous allons privilégier les modèles n'ayant besoin que d'analyses syntaxiques locales, sans représentation profonde de la syntaxe des messages.

Contexte de production

Ensuite, plus encore que pour un document textuel, un message oral ne peut être dissocié de son contexte de production (dialogue, message laissé sur un répondeur, données radiodiffusées, etc.). Comme nous avons pu le voir avec les actes de dialogues, ce contexte est indispensable pour l'obtention d'un sens *utile* à la réalisation automatisée d'un tâche.

Dimensions extra-lexicales

Enfin un des points cruciaux du traitement de la parole spontanée est qu'elle ne se résume pas à du texte lu. La transcription en mots d'un message oral ne rend compte que d'une des dimensions du message. Notamment toutes les informations prosodiques ont disparu, or celles-ci sont souvent indispensables à la compréhension de celui-ci. L'exemple le plus notoire est celui des questions exprimées avec une tournure affirmative mais dont l'intonation montante en fin de phrase suffit à classer le message comme interrogatif. De manière plus subtile la prosodie permet aussi d'ajouter une information sur le sens d'un message, par exemple lors de l'expression d'opinions ou de sentiments marqués (en positif ou négatif) ou encore lors de l'expression de l'ironie.

En dehors de la prosodie la qualité de la voix est aussi une information utile, permettant d'enrichir la description du contexte de production d'un message. Par exemple

la caractérisation en âge d'une voix (voix d'enfants, d'adultes ou de personnes âgées) peut être un paramètre important dans la réponse à donner à une intervention d'utilisateur. Cette information fait partie de la description du contexte et à ce titre participe à la notion de « sens » d'un message oral.

2.2 Quels modèles et quels algorithmes ?

Les « outils » à notre disposition pour représenter et extraire le sens d'un message oral appartiennent à deux grandes catégories : d'une part les outils d'analyse tels que ceux que l'on peut trouver pour l'analyse syntaxique et d'autre part les outils d'étiquetage et de classification qui permettent d'attribuer une étiquette à une observation donnée. Ces outils sont basés sur des modèles présentés brièvement dans les paragraphes suivants.

2.2.1 Modèles d'analyse

On peut citer parmi ces modèles les algorithmes d'analyse de type CHART de grammaires ambiguës (grammaires hors-contextes, grammaires de dépendance, ...), comme par exemple les algorithmes CYK (Younger, 1967; Kasami, 1965) et l'analyseur de Earley (Earley, 1970). Ces algorithmes, fondés sur la programmation dynamique, évitent le backtracking dans la phase d'analyse et empêchent une explosion combinatoire dans l'exploration de l'espace de recherche. Ils ont tous des complexités en $0(n^3)$, avec n étant la longueur en mot de la phrase à analyser.

Pour l'analyse de phénomènes locaux et exploitant le fait que les énoncés de parole dans la plupart des applications de dialogue homme-machine sont de courte durée, des grammaires régulières ou des approximations de grammaires hors-contexte par des grammaires régulières (Allauzen et al., 2003) sont aussi utilisées. Ces grammaires régulières sont encodées sous la forme d'automates à états finis (notés Finite State Machines ou *FSM* dans ce document). Le principal avantage de cette représentation est la possibilité d'utiliser dans la phase d'analyse l'ensemble des opérations définies sur les FSM, notamment les opérations de compositions, de déterminisation et de recherche de plus court chemin. Nous verrons dans le chapitre 4 que ce paradigme est particulièrement adapté pour lier les processus de RAP et d'analyse linguistique.

Enfin le modèle d'analyse le plus utilisé en RAP est celui des *Modèles de Langage* (ML) permettant d'attribuer une probabilité $P(W)$ à une suite de symboles $W = w_1, w_2, \ldots, w_n$. Ces modèles de langage sont généralement implémentés sous forme de Chaînes de Markov de type 2-grams (ou *bigramme*, Chaîne de Markov d'ordre 1) ou 3-grams (*trigramme*, Chaîne de Markov d'ordre 2). Si ces modèles acceptent n'importe quelle séquence de symboles comme faisant partie du langage modélisé, ils permettent d'estimer la probabilité que cette séquence a été générée par ces mêmes modèles. La recherche de la séquence W maximisant la probabilité $P(W)$ se fait avec des algorithmes efficaces tel que celui de Viterbi (Viterbi, 1967). Evidemment, ces modèles ne sont pas

exclusifs, par exemple des grammaires régulières peuvent être intégrées dans des modèles de langage, comme proposé dans (Nasr et al., 1999).

Ces modèles d'analyse nécessitent tous des corpus contenant des traces des phénomènes à modéliser. Les ML sont des modèles statistiques, ils ont besoin de corpus afin d'estimer l'ensemble de leurs paramètres (probabilités de transition entre symboles). Dans le cas des analyseurs, les grammaires sont souvent écrites de manière manuelle, néanmoins un corpus d'exemples est nécessaire afin de vérifier la couverture des règles écrites. Certaines grammaires, et notamment les grammaires modélisant des phénomènes locaux représentées par des automates tels que certaines entités nommées, peuvent aussi être induites à partir de corpus.

Par exemple, le principe de l'induction de grammaire pour la RAP est présenté dans (Vidal et al., 1995). (Wang et Acero, 2006) propose des méthodes d'induction pour le développement rapide de grammaires pour la RAP. (Dupont et al., 2005) présente de manière formelle les liens existants entre les modèles de langage de type Modèles de Markov Caché et les grammaires représentées par des automates probabilistes. La comparaison est faite autant au niveau des distributions de probabilités données par ces deux types de modèles que de leur méthodes d'apprentissage et d'inférence respectives.

2.2.2 Modèles d'étiquetage et de classification

Les outils considérés ici sont issus du domaine de l'apprentissage automatique *supervisé*. Le cadre général est le suivant : soit une suite de taille n de couples (x_i, y_i) avec $0 <= i <= n$ constituant un corpus d'apprentissage, le problème consiste à déterminer la valeur de y_j pour une valeur de x_j donnée avec $j > n$. Les termes x_i sont appelés les *observations* et les termes y_i sont les classes ou étiquettes associées à ces observations.

Deux cas peuvent se présenter :

1. soit l'observation x est une donnée atomique et dans ce cas l'attribution de la classe y ne dépend que des paramètres décrivant x ;
2. soit $x = o_1 o_2 \dots o_p$ représente une *séquence* linéaire de p observations corrélées selon un axe temporel et dans ce cas y est aussi une séquence de p étiquettes $y = e_1 e_2 \dots e_p$ où la prédiction de la classe e_j pour l'observation o_j dépend à la fois des paramètres décrivant o_j mais aussi du contexte du couple (o_j, e_j) dans la séquence (x,y).

En utilisant un cadre probabiliste le problème se résume à trouver $\hat{y}_j$ telle que :

$$\hat{y}_j = \underset{y_j}{argmax} P(y_j|x_j)$$

Généralement on effectue la distinction entre modèles probabilistes *discriminants* qui estiment directement cette probabilité $P(y_j|x_j)$, et les modèles probabilistes *génératifs* qui, eux, la renversent avec la règle de Bayes pour obtenir :

$$\hat{y}_j = \underset{y_j}{argmax} \frac{P(x_j, y_j)}{P(x_j)}$$

Dans ce cas la probabilité $P(x_j, y_j)$ permet de *générer* à la fois les observations et les étiquettes. Nous verrons dans le chapitre 4 que cette propriété est particulièrement intéressante dans le cadre de la RAP. Parmi les modèles de classification probabilistes discriminants on peut citer les modèles à base de maximisation de l'entropie (modèles *MaxEnt*) comme par exemple les modèles exponentiels de type Modèles de Markov Cachés discriminants (ou HMM MaxEnt) (Ratnaparkhi et al., 1996) et les Champs de Markov Conditionels (Conditional Random Fields - CRF) (Lafferty et al., 2001). Les modèles probabilistes génératifs les plus employés sont les Modèles de Markov Caché (ou HMM pour Hidden Markov Model).

En dehors du cadre probabiliste il existe de nombreuses méthodes de classification statistique. Le but de ces méthodes est de trouver une fonction qui minimise le risque de mauvaise classification. Parmi celles-ci on trouve les arbres de décision, les classifieurs à base de réseaux de neurones, les méthodes de régression logistique, les algorithmes fondés sur les *k-plus proches voisins*. Plus récemment des méthodes ayant pour but d'explicitement minimiser le risque d'erreur ont été développées. Les deux méthodes les plus employées sont les machines à vecteurs supports (ou *Support Vector Machine*, SVM (Vapnik, 2000)) et les algorithmes à base de *boosting* telles *Adaboost* (Schapire et Singer, 2000).

Les décisions prises par ces méthodes statistiques se traduisent généralement par un score de classification. Ces scores peuvent être utilisés pour produire par régression des probabilités que la classification soit correcte en estimant les paramètres de la régression sur un corpus d'exemples.

Enfin la plupart des méthodes de classification statistique ne fonctionnent que pour des problèmes de classification binaire. Des extensions vers des problèmes multi-classes sont néammoins possibles comme présenté dans (Allwein et al., 2001). Ceci est particulièrement important pour l'application de méthodes de classification au traitement automatique du langage.

Toutes ces méthodes, y compris les méthodes probabilistes d'étiquetage, nécessitent des corpus d'exemples pour apprendre les modèles. Le choix d'un modèle va dépendre de la taille de ce corpus d'exemples, des paramètres utilisés pour caractériser ces exemples, mais avant tout du principe choisi pour envisager la compréhension d'un message : analyse à base d'un modèle explicite du sens dans un cas ; traduction automatique d'une séquence de symboles (les mots) vers une interprétation représentée par une autre séquence de symboles (les concepts ou les structures prédicatives) dans l'autre cas. Ces deux principes sont présentés dans le paragraphe suivant.

2.2.3 Analyse syntaxique/sémantique *vs.* traduction automatique

Le traitement de transcriptions automatiques de messages oraux se caractérise par deux phénomènes particuliers : d'une part les phénomènes dus à la parole spontanée

tels que les disfluences (hésitations, reprises, auto-corrections, dislocations[3], incises) ; d'autre part l'absence de structure dans les sorties des systèmes de transcription automatique parole-texte. Cette absence de structure se caractérise par la génération d'un flux de mots sans ponctuation ni découpage en phrase. La seule segmentation généralement opérée repose sur des silences réalisés par le locuteur avec des longueurs supérieures à un seuil fixé. Ces deux phénomènes rendent très difficile toute analyse complète de ce type de message et les processus de compréhension se doivent d'opérer sur des analyses partielles. Comme présenté dans le paragraphe 2.1, l'extraction du sens d'un message peut être vue comme un processus à deux niveaux : un niveau appelé « décodage conceptuel » consistant à extraire d'un message les constituants sémantiques élémentaires (ou *concepts*) ; ce niveau est suivi d'un niveau appelé *composition sémantique* consistant à construire une structure relationnelle incluant les concepts détectés et spécifiant de manière formelle le sens du message analysé. En dehors de la traditionnelle opposition *méthodes à base de connaissances a priori / méthodes à base d'apprentissage automatique*, les méthodes développées pour répondre au problème du décodage conceptuel d'un flux de mots peuvent être vues selon deux perspectives :

- soit comme une conséquence d'un processus d'analyse (*parsing*) fondé sur un modèle sémantique (et éventuellement syntaxique) qui va construire une ou plusieurs analyses structurées du message à analyser ; dans ce cas les concepts à détecter sont des nœuds dans la ou les structures obtenues ;
- soit comme le résultat d'une opération de traduction automatique qui consiste à transformer une suite de symboles donnés en entrée (les mots) en une autre suite de symboles en sortie (les concepts). L'étape de composition sémantique est disjointe, opérant sur les concepts précédemment obtenus.

La deuxième famille de méthodes se rapproche du cadre théorique utilisé dans les applications de RAP. Dans ce cadre le décodage de parole est vu comme la transmission d'un signal dans un canal bruité. Le but est de décoder le message initial à partir des observations [4] qui ont transité à travers le canal de communication. Cette opération de *traduction* se réalise de manière probabiliste en cherchant l'interprétation $\hat{I}$ qui maximise la probabilité $P(I|A)$, A représentant la séquence d'observations acoustiques. Cette approche, initiée par les travaux de (Vidal et al., 1993) et (Levin et Pieraccini, 1995), se retrouve dans de nombreux systèmes de décodage conceptuel.

Une fois les concepts de base obtenus, le sens global d'un message peut être obtenu toujours par des méthodes fondées sur l'apprentissage automatique, telles que des classifieurs comme dans les applications de routage d'appel, par exemple How May I Help You ? (Gorin et al., 1997) ; par des méthodes à base de règles décrivant de manière explicite le processus de composition, comme dans le système Verbateam implémenté dans le service France Télécom 3000 (Damnati et al., 2007a) ; ou encore par des processus d'analyse mélangeant modèles explicites et classification automatique comme proposé dans le projet européen *LUNA* (Raymond et al., 2007).

[3] Processus linguistique de détachement d'un constituant en tête ou en fin de phrase, constituant ensuite repris par un pronom

[4] des paramètres acoustiques dans le cadre de la RAP, des mots dans le cadre du décodage conceptuel

Le choix de l'une ou l'autre de ces approches est dicté par les spécifications du système que l'on souhaite développer. Ces spécifications intègrent le niveau de robustesse voulu et la complexité de la représentation sémantique nécessaire et conditionnent à la fois le choix de l'approche mais aussi des modèles et outils utilisés pour l'implémenter.

2.2.4 Spécification d'un système de compréhension et choix d'un modèle

Les multiples approches et modèles étudiés pour répondre au problème de la compréhension de messages oraux n'ont pas permis de faire émerger un modèle « dominant » unique quelles que soient les applications visées. Tout au plus pouvons nous dégager de grandes tendances au niveau de l'utilisation des modèles et algorithmes présentés précédemment selon la complexité de l'application visée. Cette complexité peut s'exprimer selon trois dimensions : nature du langage accepté, complexité de la représentation sémantique nécessaire, prise en compte des incertitudes de la RAP. Nous les détaillons dans les paragraphes suivants.

Nature du langage accepté

Même si les messages vocaux considérés dans cette étude contiennent tous de la *langue naturelle* par opposition aux messages de commandes fondés sur des grammaires régulières simples, cette *langue naturelle* peut être caractérisée par deux paramètres qui vont conditionner la complexité de son traitement :

1. le degré de spontanéité ;
2. l'acceptation de messages ou de portions de messages hors-domaine par rapport à l'application visée.

Le degré de spontanéité permet de distinguer deux types de parole : la parole *préparée* et la parole *spontanée*. En reprenant une définition proposée par (Luzzati, 2004), cette dernière peut être définie comme étant « un énoncé conçu et perçu dans le fil de son énonciation ». Elle se caractérise par des phénomènes d'hésitation (mots tronqués et introduction de marqueurs tels que « *euh* » ou « *hum* »), de répétitions (telles que « *le le système ...* ») et d'auto-corrections (« *le lundi euh mardi après-midi* »). Ces phénomènes sont regroupés sous le terme de *disfluences*.

A cela se rajoute le phénomène de l'*agrammaticalité* : la parole spontanée se caractérise aussi par des fautes d'accord, en genre et en nombre, des phrases tronquées dont une ou plusieurs des composantes indispensables à l'analyse syntaxique sont manquantes.

Disfluences et agrammaticalité sont parmi les différences majeures existant entre textes écrits et transcriptions de l'oral. Elles sont difficiles à gérer à cause de leur caractère aléatoire. Notamment les modèles explicites de type grammaires hors-contexte échouent généralement à analyser des messages contenant de pareils phénomènes. Les modèles d'analyse statistique de type *Modèles de langage* sont, eux, plus robustes au traitement de la parole spontanée : d'une part grâce au fait que ces modèles ne rejettent rien

et sont capables d'analyser n'importe quelle suite de mots tout en gardant une fenêtre d'analyse partielle par rapport à l'ensemble du message ; d'autre part par la possibilité d'apprendre dans une certaine mesure ces phénomènes en entraînant les modèles sur des transcriptions de l'oral contenant des disfluences et des phrases tronquées.

La possibilité de traiter des énoncés ou portions d'énoncés hors-domaine est aussi une caractéristique importante des systèmes de compréhension automatique. Comme nous le verrons dans le chapitre 3 traitant des corpus utilisés dans cette étude, la notion de *hors-domaine* concerne aussi bien les commentaires que les utilisateurs peuvent faire lors de l'utilisation du système (par exemple : « alors qu'est-ce qu'il me demande déjà » ou « mais ça je l'ai déjà dit oh mais je comprends plus rien ») que les énoncés traitant réellement d'un domaine non-couvert par l'application. Là encore les modèles statistiques sont les plus robustes à ce type de parole puisqu'il n'est pas besoin de modéliser finement tous ces événements imprévus, les modèles n-grammes pouvant accepter n'importe quelle séquence de mots. De plus des modèles de rejet de type modèles *filler* peuvent être utilisés pour filtrer ces segments hors-domaine, comme nous le verrons dans le chapitre 6.

En conclusion nous pouvons dire que les systèmes fondés sur des méthodes d'analyse explicite, à partir de grammaires syntaxique/sémantique, sont adaptés au traitement de la parole *préparée*, pertinente par rapport à l'application visée (Denis et al., 2006). A l'inverse la prise en compte de disfluences, d'agrammaticalités et de segments hors-domaine, de manière robuste, nécessite le plus souvent l'utilisation de modèles statistiques.

Cette situation est similaire à celle que l'on peut trouver en RAP : alors que les systèmes grand-vocabulaire *génériques* utilisent tous des approches à base de modèles de langage statistiques de type n-grammes, les systèmes industriels de dialogue simples (petit vocabulaire, parole préparée) sont souvent implémentés à base de grammaire hors-contexte, permettant d'assurer d'un seul coup la reconnaissance et l'analyse au sein de la même phase.

Complexité de la représentation sémantique

La sous-section 2.1 a présenté plusieurs méthodes de représentation du sens d'un message. Ce choix est largement conditionné par la complexité de la tâche envisagée par le système utilisant le module de compréhension. Ces tâches sont de deux sortes, illustrant les deux grands domaines d'utilisation de la langue naturelle :

1. le langage comme outil de communication : ce domaine regroupe l'ensemble des systèmes de dialogue homme-machine ;
2. le langage comme outil de représentation de connaissances : on trouve ici les systèmes d'extraction d'informations à partir de documents audio.

Dans chacun de ces domaines se trouve une très grande variété d'applications. Ainsi les systèmes de dialogue peuvent être catégorisés en trois types d'applications de complexité croissante : les systèmes de routage d'appel, les systèmes de type « remplissage

de formulaire », et enfin les systèmes nécessitant une négociation entre la machine et les utilisateurs. Si les deux premiers types de systèmes peuvent se contenter de représentations sémantiques assez simples, très proches de l'application, comme celles de *How May I Help You ?* (Gorin et al., 1997) ou de *FT3000* (Damnati et al., 2007a), les systèmes nécessitant une négociation se doivent de construire une représentation structurée prenant en compte l'ensemble du dialogue, créant des liens entre les entités reconnues à chaque tour de parole et entre les tours.

De la même manière les systèmes d'extraction d'information se distinguent par la complexité des informations recherchées : de la recherche simple d'entités, de type entités-nommées comme nous le présenterons dans le chapitre 4, à la caractérisation de messages comme pour la détection d'opinions présentée dans le chapitre 6, jusqu'à l'extraction d'entités structurées comme dans la tâche de *distillation* (Hakkani-Tur et Tur, 2007) d'information du récent programme de recherche américain *GALES* (Olive, 2005).

Le choix de la représentation sémantique va directement dépendre de l'application visée : dans les deux domaines présentés, toutes les applications nécessitant uniquement l'extraction d'entités sans nécessairement construire de liens entre elles peuvent utiliser des représentations « à plat » de type *attribut-valeur*. Par contre, dès qu'une structure est nécessaire, des modèles intégrant la notion de *prédicat* sont indispensables.

Prise en compte des incertitudes de la Reconnaissance Automatique de la Parole

Nous verrons dans le chapitre 4 un modèle permettant d'intégrer les processus de reconnaissance et de compréhension de la parole. Cette intégration, ou tout du moins cette collaboration, entre ces deux processus est aussi un des paramètres influençant le choix des modèles utilisés pour la compréhension.

La plupart des systèmes de compréhension de l'oral sont fondés sur une approche séquentielle où les phases de transcription et de compréhension sont dissociées : les modèles de RAP produisent d'abord une séquence de mots $\hat{W}$ sur laquelle sont appliqués les processus de compréhension tels que présentés au paragraphe précédent. Cette approche à l'avantage de pouvoir utiliser directement tous les outils développés en Traitement Automatique de la Langue Naturelle (TALN) pour la langue écrite (outils d'analyse ou de classification). Elle est cependant sous-optimale, comme présenté dans (Wang et al., 2005), dans la mesure où le lien de dépendance entre le signal de parole (représenté par les observations acoustiques A) et l'interprétation du message $\hat{I}$ est rompu. Aucune information liée au contenu sémantique du message n'est utilisée pour obtenir $\hat{W}$, estimé par le maximum de la probabilité a posteriori : $\hat{W} = \underset{W}{argmax} P(W|A)$.

Dans la deuxième approche les deux processus collaborent par la production non pas d'une hypothèse unique par le module de RAP mais d'un ensemble d'hypothèses valuées sous la forme d'une liste de *n-meilleures* solutions ou bien sous la forme d'un graphe d'hypothèses. Le choix final d'une hypothèse dépend à la fois des scores ou mesures de confiance attachés à chaque hypothèse et aussi des scores donnés à chacune

d'entre elles par le module de compréhension. Garder un espace de recherche ouvert est important car les chaînes de mots données en sortie des systèmes de RAP sont produites à l'aide de modèles de langage ayant une portée très faible (modèles bigrammes ou trigrammes), ne garantissant aucune cohérence au delà d'une fenêtre de quelques mots.

L'avantage d'une telle approche est également la possibilité de retarder la décision sur les choix des mots reconnus en utilisant des modules de plus haut niveau que les simples modèles de langage n-grammes. Par exemple on peut utiliser le contexte de production d'un message pour désambiguïser les hypothèses produites et augmenter ainsi la robustesse du système. Cependant cette approche nécessite une modification profonde de nombreux modèles développés sur la langue écrite afin de pouvoir prendre en compte à la fois un ensemble d'hypothèses multiples, et surtout une incertitude sur les mots reconnus représentée par les scores attachés à chacun d'eux.

Ainsi, si les modèles d'analyse de type Chaîne de Markov ou d'étiquetage comme les Modèles de Markov Cachés (ou HMM pour Hidden Markov Model) se combinent naturellement avec les sorties d'un module de RAP dans la mesure où le paradigme probabiliste des deux processus est similaire, il n'en va pas de même pour les outils d'analyse à base de grammaire demandant d'importantes modifications pour pouvoir prendre en compte des graphes d'hypothèses valuées.

Par exemple plusieurs méthodes ont été proposées afin d'appliquer des analyseurs syntaxiques à des graphes de mots. Certaines de ces méthodes ((Chappelier et al., 1999; Kieffer et al., 2000)) ont pour but de produire efficacement des arbres d'analyse avec des algorithmes de type *chart*, sans réévaluer les arcs des graphes. D'autres méthodes emploient des grammaires afin de calculer le meilleur chemin dans le graphe, au moyen d'un algorithme de type A$*$ (Chelba et Jelinek, 2000) ou d'un analyseur Markovien modifiant les poids de chaque transition (Roark, 2002). Une autre solution consiste à employer des grammaires régulières pouvant facilement être représentées par des machines à états finis (FSM). L'application de FSMs à des graphes de mots est une opération parfaitement définie car ces mêmes graphes de mots peuvent également être représentés par des FSMs (voir (Mohri et al., 2000, 2002) pour plus de détails sur les grammaires régulières et les FSMs).

2.2.5 Discussion

Je vais présenter maintenant quelques approches utilisées pour construire des systèmes de compréhension à partir des outils et des spécifications présentées précédemment.

Approches à base de grammaires

Les approches à base de grammaires ont pour but de donner une valeur de vérité à un symbole non-terminal étant donné la présence d'autres symboles terminaux et

non-terminaux. Des symboles représentant des entités sémantiques peuvent être ajoutés à ces grammaires en complément de ceux définissant des structures syntaxiques. On peut ainsi obtenir, à l'issue d'une phase d'analyse syntaxique, des hypothèses sémantiques à partir d'une séquence de mots issus d'un module de RAP. Même si idéalement ces grammaires devraient être dépendantes du contexte puisque l'interprétation d'un message oral se fait toujours par rapport à son contexte d'élocution, on utilise généralement des grammaires hors contextes pour des contraintes d'efficacité. Afin de prendre en compte l'ambiguïté d'analyse et l'imprécision dans la séquence de mots issus de la RAP, ces grammaires peuvent être rendues stochastiques en utilisant un corpus d'apprentissage.

Quand l'analyse complète échoue, à cause par exemple de disfluences ou d'erreurs de reconnaissance, les analyseurs CHART produisent des forêts de sous-arbres représentant un ensemble d'analyses partielles. Le problème portant sur l'estimation de la probabilité d'une analyse partielle avec une grammaire hors-contexte probabiliste a été étudié dans (Corazza et al., 1994).

Pour illustrer cette famille de méthodes on peut citer le système du MIT TINA (Seneff, 1992b) fondé sur des analyses syntaxico-sémantiques complètes et partielles. Ce système utilise un ensemble de règles hors-contexte probabilistes de réécriture avec contraintes, qui est converti automatiquement sous forme de réseau dans lequel chaque nœud représente une catégorie syntaxique ou sémantique (Seneff, 1992b). Les probabilités liées aux règles sont estimées sur un corpus d'apprentissage. Elles servent à contraindre la recherche pendant l'analyse. Afin d'augmenter la robustesse du système, des analyses partielles sont possibles (Seneff, 1992a). Dans ce cas l'analyseur produit, pour chaque mot du message, l'ensemble des analyses pouvant commencer à cet endroit.

D'autres systèmes utilisent une analyse robuste incrémentale fondée sur une étape de *chunking* comme dans (Antoine et al., 2003) ; (Wang et al., 2002) présente des grammaires hors-contexte syntaxico-sémantiques dérivées à partir de patrons génériques. Dans la campagne d'évaluation des systèmes de compréhension de la parole Technolangue-Evalda-MEDIA, dont ce document présente le système développé au LIA dans le chapitre 4, cette approche était représentée par les systèmes du VALORIA et du LORIA (Denis et al., 2006).

Dans le travail décrit dans ce document les seules grammaires qui vont être utilisées sont des grammaires régulières codées sous la forme d'automates à états finis. La justification de ces grammaires est de modéliser des phénomènes très locaux, de taille limitée, constituant les « briques » de base du module d'interprétation comme présentés dans le chapitre 4.

Analyse linguistique robuste

Les approches fondées uniquement sur des grammaires ont généralement des problèmes de couverture et de robustesse face aux phénomènes de l'oral spontané et des erreurs de RAP. Afin d'augmenter la robustesse de l'analyse, il a été proposé des modèles

d'analyse syntaxique s'intégrant dans le processus de RAP, fondés sur des modèles de langage robustes incluant des contraintes syntaxiques. Par exemple, le Structured Language Model (SLM) proposé par (Chelba et Jelinek, 2000), et étendu à l'analyse sémantique dans (Bod, 2000).

Une autre approche, fondée sur le concept d'analyse locale développée dans (Abney, 1998), consiste à n'utiliser que des contraintes syntaxiques locales pour extraire les composants sémantiques nécessaires à la compréhension d'un message. La composition de ces concepts en une interprétation est effectuée par un autre niveau où n'interviennent plus de contraintes syntaxiques. C'est une approche de ce type que nous avons utilisée pour développer le système de compréhension sur le corpus MEDIA présenté dans le chapitre 4.

Enfin une approche plus récente applique cette notion *d'étiquetage de surface* ou « *shallow parsing* » à l'analyse sémantique en proposant des stratégies de « *semantic shallow parsing* » consistant à détecter des *rôles sémantiques* dans un énoncé, ces rôles constituant une représentation sémantique de haut niveau indépendante d'une application en particulier. Cette approche a été initiée par les travaux de (Gildea et Jurafsky, 2002) et (Pradhan et al., 2004). Le texte est tout d'abord analysé à l'aide d'analyseurs syntaxiques de surface puis des classifieurs prenant de nombreux paramètres en entrée vont prédire les rôles sémantiques présents dans la phrase qui vont permettre de construire une structure à base de prédicats représentant l'interprétation de la phrase.

La compréhension comme une tâche de classification

Lorsqu'une représentation sémantique *à plat* est suffisante pour l'application visée, de nombreux systèmes ramènent le problème de la compréhension à une tâche d'étiquetage ou de classification. Parmi les premiers systèmes implémentant cette approche on peut citer le système Chronus développé pour la tâche ATIS et utilisant des arbres de classification sémantiques (ou Semantic Classification Trees - SCT) (Kuhn et De Mori, 1995). Depuis de très nombreux systèmes suivent cette approche, par exemple des systèmes de routage d'appel comme le système d'AT&T *How May I Help You ?* (Gorin et al., 1997) qui utilise des classifieurs à base de SVM ou de Boosting (Haffner et al., 2003).

Enfin, suivant les travaux de (Vidal et al., 1993) et (Levin et Pieraccini, 1995), le problème de la compréhension peut être ramené à un problème de classification ou d'étiquetage de séquences, le but étant d'associer à une séquence de mots, une séquence de concepts modélisant le sens. Par exemple, toujours dans la campagne MEDIA les deux systèmes du LIMSI (Bonneau-Maynard et Lefevre, 2005) et celui du LIA (Servan et Bechet, 2006) implémentent cette approche. Différents modèles d'étiquetage de séquences peuvent être utilisés, on trouvera dans (Raymond et Riccardi, 2007) une comparaison de différentes méthodes d'étiquetage à base de HMM, de CRF et de SVM, toutes comparées sur deux corpus : ATIS pour l'anglais et MEDIA pour le français. Des réseaux bayésiens ont été utilisés, par exemple dans (Jamoussi et al., 2003), combinés avec des CRF dans (Lefevre, 2006). L'association entre des séquences de mots et des commandes d'un système de dialogue est obtenue par l'analyse sémantique latente (*Latent Semantic*

Analysis - LSA) dans (Bellegarda et Silverman, 2000).

2.3 Quels corpus et quelles observations pour l'apprentissage des modèles ?

L'essentiel des méthodes utilisées dans cette étude sont des méthodes fondées sur l'apprentissage (automatique ou manuel) à partir de *corpus*. Dans ce cadre le terme *corpus* désigne une collection d'*observations* relatives aux phénomènes linguistiques que l'on souhaite modéliser. De même que les modèles présentés dans la section 2.2 se divisent en deux catégories (modèles d'analyse et modèles de classification/étiquetage), les *observations* contenues dans ces corpus sont de deux types :

1. pour les modèles d'analyse les observations constituent l'ensemble des phrases acceptables, avec la fréquence de chacune d'elle :
2. pour les modèles de classification ou d'étiquetage les observations sont des couples (x_i, y_i) où l'événement x_i est associé à la classe ou l'étiquette y_i.

Par exemple si l'on dispose d'un corpus contenant des transcriptions manuelles de messages oraux dans lesquelles les entités *dates* sont marquées, on peut :

- soit extraire de ce corpus uniquement les séquences de mots produisant une date et utiliser ces exemples pour construire un modèle reconnaissant ces entités (grammaire manuelle ou induite des données, grammaire simple ou stochastique, modèle de langage statistique) ;
- soit construire un corpus d'exemples (les séquences représentant les dates) et de contre-exemples (toutes les autres séquences) et entraîner un classifieur à déterminer automatiquement ce qui sépare ces deux types de séquences de mots.

En dehors de cette opposition sur le paradigme utilisé pour envisager le problème de la compréhension de messages, les corpus utilisés se distinguent les uns des autres par deux caractéristiques : les conditions de collecte des messages et la nature des observations utilisées pour décrire chaque message. Ces caractéristiques sont discutées dans les deux prochaines sous-sections.

2.3.1 Conditions de collecte

En fonction des conditions de collecte on distingue trois types de corpus :

1. les corpus *écologiques*, c'est à dire collectés dans des situations réelles, non destinées à l'étude, par exemple les corpus d'émissions diffusées (*Broadcast News*) ou les corpus collectés dans les centres d'appel avec opérateurs humains ;
2. les corpus *de laboratoire*, collectés dans le cadre d'études scientifiques où les locuteurs, bénévoles ou rémunérés, savent qu'ils participent à une étude et ont un protocole à suivre ;
3. enfin une dernière génération de corpus a vu le jour récemment, il s'agit de corpus homme-machine collectés dans le cadre d'applications de dialogue homme-machine déployées auprès de vrais clients.

Les corpus écologiques

Les corpus *écologiques* contiennent des données *réalistes*, c'est à dire conformes à ce que doit traiter un système automatisé destiné à un large usage et à des utilisateurs novices, notamment les phénomènes liés à la parole spontanée et les énoncés hors-domaines sont très présents. La quantité de corpus disponible est potentiellement infinie. En effet la masse de documents audio disponible, soit enregistrés et diffusés par radio, télévision et site internet, soit accessibles en enregistrant les conversations entre des opérateurs et des utilisateurs dans des centres d'appel, est en perpétuelle augmentation. De plus le faible coût de son stockage rend possible l'exploitation de ces données sur une large échelle.

Les inconvénients de ce type de collecte résident d'abord dans la complexité des messages collectés. Par exemple les conversations homme-homme que l'on trouve dans les centres d'appel ne sont pas similaires aux conversations homme-machine que l'on désire modéliser (Riccardi et Gorin, 1998) lors du développement d'un serveur vocal automatisé. Or souvent, avant le développement de tout système, ce sont les seules données disponibles pour amorcer les processus d'apprentissage. De plus en l'absence de contexte de collecte contrôlé et en raison du degré de spontanéité des données collectées, il est difficile d'appliquer des systèmes automatiques d'annotation sur les messages audio. Par conséquent l'effort manuel, et donc le coût, que nécessite l'annotation de tels corpus est important.

Ce point est particulièrement vrai pour les corpus collectés dans les centres d'appel. Les corpus représentant des émissions diffusées contiennent principalement deux types de messages de complexité très différentes : les messages correspondant à de la parole *préparée*, et ceux contenant de la parole *spontanée*. Pour la parole préparée les locuteurs sont souvent des locuteurs professionnels (animateur ou homme politique par exemple). Elle est énoncée dans un environnement contrôlé (studio d'enregistrement) et elle s'apparente à de la parole lue. Les messages contenant de la parole spontanée, souvent avec plusieurs locuteurs (débats ou interviews), dans des conditions acoustiques variables (reportage à l'extérieur, communication téléphonique), posent de très nombreux problèmes de segmentation.

Pour le premier type de messages il est assez facile de se procurer des données d'apprentissage, notamment la langue utilisée est proche de la langue écrite telle qu'on peut la trouver dans des corpus de textes journalistiques disponibles en très grande quantité. De plus les conditions acoustiques excellentes d'enregistrement rendent ces données faciles à traiter par un processus d'analyse automatique. Par contre pour le deuxième type de messages, les difficultés de segmentation (séparation bruit/musique/parole, segmentation et suivi de locuteur) ainsi que les difficultés intrinsèques à l'oral spontané (disfluences, variabilité lexicale et syntaxique) rendent ces messages très difficiles à traiter de manière automatique. Dans ce document le corpus ESTER pour l'extraction d'entités nommées et le corpus de sondage d'opinions de France Télécom font partie de ces corpus *écologiques*.

Les corpus de laboratoire

Le deuxième type de corpus est obtenu en laboratoire, suivant un protocole précis, dans un but d'étude ou de prototypage d'un système de compréhension. Les locuteurs sont soit bénévoles soit rémunérés. Ils suivent des instructions leur décrivant le rôle qu'ils doivent jouer durant la collecte ainsi que d'éventuelles consignes sur le type de réactions qu'ils doivent avoir face à la machine ou à l'opérateur simulant une machine. Ces protocoles de collecte sont indispensables au développement de chaque nouvelle application de dialogue, cependant ils sont fondés sur deux paradoxes :

1. Le premier paradoxe concerne la disponibilité d'un système permettant cette collecte. En effet le but est de collecter des interactions aussi réalistes que possible entre des utilisateurs et le système que l'on cherche à mettre au point. Ces interactions peuvent être vues comme une anticipation de ce que devra être le système. Le but des processus d'apprentissage est de modéliser et de reproduire cette anticipation. Or si le système utilisé lors de la collecte est trop déficient, les interactions collectées ne pourront servir à améliorer celui-ci puisqu'elles ne feront que mettre en avant ces déficiences. C'est pour cette raison qu'on utilise souvent un processus de *simulation* pour le système de collecte fondé sur le paradigme du *Magicien d'Oz*, proposé dans (Gould et al., 1983), baptisé *Wizard of Oz* ou *WOZ* dans (Kelley, 1984).

2. Le deuxième paradoxe est relatif au « réalisme » des interactions obtenues. Puisqu'il s'agit d'expériences de laboratoire, les sujets se prêtant à la collecte ont nécessairement des comportements différents de ceux que peuvent avoir de vrais utilisateurs. Ces différences se situent à la fois dans leurs réactions face à la machine mais aussi dans le langage employé. Concernant la gestion du dialogue, comme il s'agit d'expérimentations contrôlées, les sujets ont généralement un scénario et des consignes à suivre. Si ce scénario est trop explicite les sujets auront tendance à ne faire que reproduire le scénario proposé, limitant la quantité d'informations nouvelles que peut contenir le corpus. Sur le langage employé, là aussi il faut faire attention à ne pas trop lexicaliser les instructions données aux sujets, le risque étant que ceux-ci se contentent de lire les instructions, ruinant par la même l'intérêt du corpus collecté. Toute la difficulté de ce type de collecte réside dans le difficile équilibre entre les consignes données aux sujets pour contrôler l'expérimentation et la nécessaire liberté qu'ils doivent avoir afin de produire des interactions les plus réalistes possibles.

De nombreux protocoles ont été proposés pour la simulation par *Magicien d'Oz*. Par exemple pour les systèmes de dialogue du MIT (Glass et al., 2000) ou de CMU (Eskenazi et al., 1999). Les sujets ne savent pas forcément que le système avec lequel ils interagissent est une simulation. Des études récentes proposent même de placer des protocoles de type *Magicien d'Oz* dans des centres d'appel, permettant de collecter ainsi des interactions auprès de sujets ignorant qu'ils participent à une expérience, par exemple le *ghost wizard* proposé par AT&T (Fabbrizio et al., 2005), ou les travaux reportés dans (Wiren et al., 2007).

Etant donné la difficulté de la mise au point du protocole et de la collecte des données, il existe peu de corpus de laboratoire contenant des dialogues homme-machine contrôlés et disponibles auprès de la communauté scientifique par l'intermédiaire des agences de distribution de corpus telles que LDC aux Etats-Unis ou ELDA en Europe. Les deux exceptions notoires sont le corpus ATIS (Pallett et al., 1992) pour l'anglais et le corpus MEDIA (Bonneau-Maynard et al., 2005) pour le français.

Du point de vue de l'annotation, le nombre restreint de messages susceptibles d'être obtenus à cause du coût de recrutement des sujets, le caractère *contrôlé* de la collecte ainsi que la part limitée des phénomènes dus à la parole spontanée rendent ces corpus relativement faciles à annoter. Notamment, une des principales différences entre ces messages et ceux des corpus *écologiques* est l'absence d'énoncés hors-domaine.

Dans ce document les corpus de laboratoire étudiés sont le corpus PlanResto (Sadek et al., 1996) de France Télécom et le corpus MEDIA (Bonneau-Maynard et al., 2005).

Les corpus provenant d'applications mises en service

Enfin le dernier type de corpus correspond aux fichiers *log* (ou *trace*) obtenus à partir d'applications automatisées, prenant en entrée de la parole naturelle, mises en service sur une grande échelle auprès de *vrais* utilisateurs. Ces applications se sont déployées au début des années 2000, notamment l'application d'AT&T *How May I Help You ?* (Gorin et al., 1997), qui a été sans doute l'application la plus largement déployée à ce jour, traitant quotidiennement une importante partie des appels clientèles de la compagnie à l'échelle des Etats-Unis.

Les caractéristiques de ces corpus présentent beaucoup de points positifs :
- vrais utilisateurs, ayant un réel intérêt à la complétion de la tâche, contrairement aux corpus de laboratoire ;
- parole spontanée, large variabilité de locuteurs, d'accents et de comportements :
- situation d'interaction homme-machine, pas de trace de communication non-verbale comme on peut en trouver dans des corpus d'interactions homme-homme ;
- annotations automatiques déjà faites par le système, avec éventuellement corrections manuelles obtenues lorsque le système s'est trouvé dans une impasse et qu'un opérateur humain a pris le relais ;
- enfin la quantité de données est potentiellement illimitée tant que le système est mis en service.

Cette succession de points positifs a cependant son revers.

Tout d'abord ces corpus sont difficilement accessibles. En effet, en raison de leur aspect stratégique au niveau commercial, ils sont la propriété des opérateurs ayant déployé les services à partir desquels ils ont été obtenus. De plus certains d'entre eux peuvent contenir des informations personnelles sur les utilisateurs, ce qui empêche leur diffusion publique. Ils ne sont accessibles à la recherche académique que dans le cadre de partenariat encadrant précisément leur utilisation. Une des particularités des travaux présentés dans cette habilitation à diriger des recherches est d'avoir pu accéder à de tels corpus à travers des collaborations étroites avec deux importants opérateurs

télécoms : AT&T aux Etats-Unis sur le corpus *How May I Help You ?* et France Télécom en France sur le corpus *FT3000*.

Le deuxième inconvénient de ce genre de corpus est le fait que ces fichiers log constituent une « photo » du système à un instant donné. Les interactions du système avec des utilisateurs sont forcément dépendantes des performances de celui-ci. Notamment les contraintes de robustesse font que les fonctionnalités implémentées sont généralement assez restreintes, conduisant à des interactions elles-mêmes peu complexes. Par exemple la durée moyenne, en nombre de tours de parole, des dialogues des corpus *How May I Help You ?* et *FT3000* n'excède pas 3 tours de parole.

On trouve ici un nouveau paradoxe : comment utiliser des traces de systèmes volontairement peu complexes à cause de contraintes de robustesse afin de développer une *nouvelle génération* de systèmes, à même d'appréhender des tâches plus ambitieuses, tout en conservant la robustesse nécessaire. Un exemple du passage d'un système à base de règles vers un système stochastique obtenu grâce à de tels corpus est présenté dans le chapitre 5.

2.3.2 Nature des observations annotées

La nature des observations annotées dans les corpus d'apprentissage est cruciale car la vocation des modèles utilisés dans cette étude est de modéliser et reproduire ces annotations sur de nouvelles données. Ces observations dépendent directement du modèle sémantique utilisé pour représenter le sens d'un message, comme présenté dans le paragraphe 2.1. Ainsi le type des observations va différer selon leur proximité avec le cadre applicatif dans lequel les corpus ont été collectés. D'une manière générale plus une annotation est proche de l'application, plus elle est directement utilisable pour construire un modèle opérationnel. Cependant elle demandera aussi plus d'efforts pour les annotateurs qu'un modèle d'annotation générique pouvant se baser sur des outils de Traitement Automatique de la Langue tels que des étiqueteurs morphosyntaxique ou des analyseurs syntaxique.

Par exemple, pour une application de routage d'appel, une annotation où chaque message est étiqueté avec la destination prévue dans le système a le mérite de pouvoir être utilisée directement pour entraîner un processus de classification supervisé. Cependant cette annotation ne pourra être faite que par un expert du domaine visé connaissant parfaitement le système. De plus toute modification dans la structure de l'application de routage nécessite une réannotation du corpus. A l'inverse si l'annotation se fait sur des notions plus génériques telles que les *entités nommées* et les actes de dialogue, celles-ci sont indépendantes de l'application de routage. Elles sont donc robustes à un changement dans l'application et ne nécessitent pas forcément un expert du domaine pour être posées. Par contre ces observations ne permettent pas d'entraîner directement un classifieur d'appel, il faut rajouter une étape faisant le lien entre ces annotations et la destination des appels.

Nous allons maintenant présenter brièvement les différents types d'annotations que l'on peut trouver dans les corpus oraux à la base de ce travail.

Segmentation du signal

Les premières annotations nécessaires au traitement d'un corpus de message oraux sont les informations concernant la segmentation du signal. Tout d'abord la segmentation en type de signal, par exemple : bruit, musique, parole, parole sur musique, téléphone ; puis la segmentation en locuteurs, avec si possible le suivi des locuteurs dans le cas de messages contenant plusieurs d'entre eux tels que des interviews, des débats ou des conversations. Enfin la segmentation en groupes de souffle. En effet les messages oraux ne contiennent évidemment aucune ponctuation permettant de les structurer. Cette structuration peut être faite avec des critères prosodiques, cependant en raison de la difficulté d'étiqueter prosodiquement un corpus, les seules indications généralement marquées sont celles des groupes de souffle.

Transcription en mots

La transcription en mots des corpus oraux est un préalable aux annotations de plus haut niveau. Des conventions sont généralement adoptées pour traiter les cas ambigus de transcription tels que les mots tronqués ou les disfluences, comme par exemple celles définies dans le cas du projet *Transcriber* (Barras et al., 2001). Dans certains cas il peut être intéressant d'ajouter des informations relatives à la prononciation en plus de la transcription. Par exemple, l'annotateur peut préciser le mode de prononciation d'un acronyme : lu ou épelé. Il peut aussi indiquer qu'un mot a été prononcé en ne respectant pas les règles phonologiques usuelles, ou bien encore qu'il s'agit d'un mot d'emprunt à une langue étrangère, en particulier les noms propres. Ces informations sont importantes si le corpus doit être utilisé pour entraîner des modèles acoustiques. Cependant l'ajout de toute information supplémentaire à un impact important sur le coût humain des transcriptions.

Annotations syntaxiques

On peut trouver deux types d'annotations syntaxiques :
- les annotations « à plat », correspondant à l'étiquetage syntaxique de surface, telles que l'annotation en étiquettes morpho-syntaxiques ou *Part Of Speech* (*POS*) ainsi que l'annotation en syntagmes minimaux (ou *chunks*)
- les annotations *profondes* faisant intervenir une analyse syntaxique sur les transcriptions en mots des corpus oraux.

Les premières annotations sont généralement posées par des méthodes d'étiquetage automatique du fait de la bonne robustesse des systèmes d'étiquetage morpho-syntaxique et de découpage en syntagmes minimaux.

En ce qui concerne les annotations syntaxiques profondes, de type arbres d'analyses comme dans le corpus *PennTreeBank* (Marcus et al., 1994), il n'existe pas d'exemples, à notre connaissance, d'autres corpus de grande taille contenant de telles annotations manuelles pour des transcriptions de l'oral. Un effort dans cette direction a été fait pour

la campagne d'évaluation des analyseurs syntaxiques Technolangue EASY (Paroubek et al., 2005) et les difficultés d'annotation syntaxiques de corpus oraux sont discutées dans (Benzitoun et Veronis, 2005).

Annotation en concepts

Le terme « concept » est volontairement ici très vague. Il concerne toutes les annotations sémantiques « à plat » où une séquence de mots est associée à une paire attribut/valeur. Ces paires peuvent représenter des entités nommées, basiques comme définies dans MUC7 (Chinchor et Robinson, 1998) ou très détaillées comme dans (Sekine et al., 2002) ; des *Frame Element* comme dans le projet *FrameNet* ; des unités lexicales telles que mots-clés ou séquences de mots-clés représentant des *briques* de sens pour une application donnée. Ces concepts peuvent être aussi des unités *ad hoc* définies pour un système en particulier et sans vocation générique.

Il existe très peu de corpus oraux disponibles ayant un tel étiquetage, à l'exception des corpus d'émissions diffusées étiquetés en entités nommées, tel que le corpus ESTER. Concernant les corpus de dialogue, le seul corpus de grande taille possédant un tel étiquetage, à notre connaissance, est le corpus MEDIA qui sera présenté dans le chapitre suivant.

Annotation sémantique structurée

Ce niveau concerne l'annotation en prédicat, pouvant être obtenue en composant les concepts présentés dans le paragraphe précédent. Cette annotation peut être très proche de l'application comme dans le corpus de dialogue ATIS où chaque message est annoté avec la requête SQL lui correspondant, ou encore avoir une volonté générique comme l'annotation en prédicat définie dans FrameNet. Cependant il n'existe pas de corpus oral annoté grâce à un tel formalisme. Le projet Européen LUNA (Raymond et al., 2007) qui est à la base de plusieurs des travaux présentés dans ce document a vocation à combler cette lacune.

Annotation spécifique au dialogue oral

Les corpus de dialogue oraux posent un problème particulier pour les paradigmes d'annotation : la prise en compte du contexte du dialogue. Celui-ci a une influence d'une part sur les actes de dialogue tels que présentés dans le paragraphe 2.1 et d'autre part sur le type des concepts détectés dans un tour de parole. Par exemple un énoncé ne contenant que l'expression d'une valeur numérique ne peut être annoté, hors contexte du dialogue, que comme un élément numérique. En rajoutant le contexte cette valeur peut représenter une quantité (nombre d'items ou prix par exemple), une date, un code, un nom propre, *etc*. Par exemple le corpus MEDIA contient une double annotation :

– hors contexte : les tours de parole sont présentés aux annotateurs dans un ordre aléatoire, aucune connaissance du contexte du dialogue ne doit être utilisée ;

– en contexte : pour cette phase les dialogues sont présentés aux annotateurs dans leur ensemble, les ambiguïtés sur le type des concepts sont levées grâce au contexte du dialogue, les liens référentiels entre les différents tours sont résolus.

Sur le corpus MEDIA environ 20% des spécifieurs de concepts sont modifiés par l'introduction du contexte du dialogue comme nous le verrons dans le chapitre 3.

2.3.3 Discussion

En conclusion de cette présentation des différents types de corpus et d'annotations pouvant être utilisés pour l'apprentissage de modèles de compréhension, deux points vont être discutés : l'utilisation des corpus pour l'apprentissage automatique et l'extraction de connaissances explicites d'une part ; la granularité des annotations d'autre part.

Apprentissage automatique et extraction de connaissances explicites

Les méthodes à base d'apprentissage automatique sur corpus sont souvent opposées aux méthodes à base de représentation explicite de connaissances de la manière suivante :

Les méthodes « automatiques » nécessitent de disposer d'importants corpus d'apprentissage difficiles à obtenir mais le niveau d'expertise nécessaire à l'adaptation des modèles à une nouvelle tâche est limité.

A l'inverse, les méthodes « explicites » nécessitent un niveau d'expertise linguistique important pour construire un modèle, posant donc des difficultés d'adaptation rapide à une nouvelle tâche. Par contre aucune collecte de corpus n'est nécessaire.

Cette opposition, très schématique, est pertinente pour certaines tâches du Traitement Automatique des Langues telles que l'étiquetage morpho-syntaxique ou la détection d'entités nommées génériques à partir de corpus de texte écrit. En effet dans ces deux cas les méthodes automatiques obtiennent des résultats corrects pour peu que la quantité de données d'apprentissage soit suffisante. Dans le même temps les méthodes à base de connaissances peuvent obtenir des performances supérieures en apportant suffisamment de soin à la mise au point des modèles. Ces résultats ont été montrés dans plusieurs études comparatives, tels que (Samuelsson et Voutilainen, 1997) pour les analyseurs morphosyntaxiques ou la campagne d'évaluation MUC-7 (Chinchor et Robinson, 1998) pour les entités nommées.

Dans le cas des méthodes de compréhension de messages oraux la situation est moins claire. En effet en l'absence d'un modèle général de représentation du sens, chaque cadre applicatif nécessite une étude sur corpus pour déterminer la modélisation adéquate. Ainsi même les modèles à base de connaissances explicites sont soumis à une phase d'acquisition de corpus afin de collecter suffisamment d'exemples pour construire les modèles. De manière similaire les systèmes d'apprentissage automatique ont besoin, sur les corpus collectés, d'annotations sémantiques qui ne peuvent

être posées que par rapport à un modèle formel comme nous l'avons vu au début de ce chapitre. Or ce modèle d'annotation se doit d'être formalisé par des experts sous la forme d'un *manuel d'annotation* afin de permettre aux annotateurs humains d'être consistants entre eux et au fur et à mesure de l'annotation. Ce phénomène est illustré par la figure 2.1 qui présente une mesure du taux d'accord inter-annotateurs (mesure Kappa (Carletta, 1996)) en fonction de l'avancée de la collecte et de l'annotation du corpus de dialogue MEDIA. Chaque évaluation, appelée IAg, était faite après la collecte et l'étiquetage de plusieurs dizaines de dialogues. On considère généralement que le taux Kappa est satisfaisant s'il dépasse les 80%. Ce chiffre est obtenu à partir de l'IAg 4 qui a nécessité l'annotation de plus de 200 dialogues.

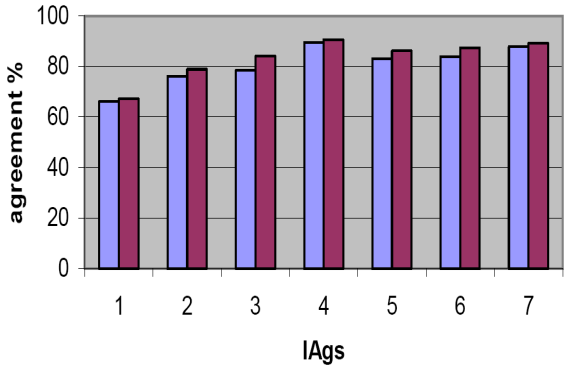


FIG. 2.1: *Variation de l'accord inter-annotateur (mesure Kappa) en fonction des différentes IAg. Les deux mesures données pour chaque IAg sont l'accord sur les séquences de concepts attributs-valeurs pour le corpus MEDIA, dans la première valeur quatre modes sont considérés pour caractériser les concepts, pour la deuxième valeur l'ambiguïté est réduite en ne considérant que deux modes*

On peut en déduire que pour la tâche MEDIA (tâche d'annotation relativement simple car représentation « à plat » du sens comme nous le verrons dans le chapitre 3), le manuel d'annotation a nécessité le traitement de 200 dialogues avant d'être suffisamment de bonne qualité pour permettre d'obtenir des annotations consistantes. La qualité du manuel d'annotation, et donc de l'expertise linguistique qui a permis de l'obtenir, est cruciale pour les méthodes d'apprentissage automatique car toute inconsistance dans l'annotation induit un biais dans l'apprentissage des modèles.

On peut donc voir une convergence des méthodes « automatiques » et « à base de connaissances » sur ces deux points : la nécessité de collecter un corpus d'exemples d'une part et la nécessaire expertise linguistique sur ce corpus d'autre part.

Pour les méthodes fondées sur la modélisation explicite de connaissances ce corpus permet d'obtenir un modèle procédural (grammaires ou bases de règles) développé sur les exemples collectés.

Pour les méthodes d'apprentissage automatique sur corpus l'expertise linguistique permet d'obtenir un modèle descriptif définissant le manuel d'annotation du corpus.

Un autre point justifiant la collecte systématique de corpus, quelles que soient la méthode employée, est le fait que l'analyse de messages oraux nécessite l'utilisation de systèmes de transcription automatique de parole. Or le traitement de la parole spontanée implique l'utilisation systématique, de la part de ces systèmes, de modèles de langage statistiques de type *n-grammes* qui doivent être entraînés sur des corpus contenant des transcriptions manuelles aussi proches que possibles des messages qui doivent être traités. Plus le corpus d'apprentissage des modèles *n-grammes* est proche du cadre applicatif visé, meilleures seront les transcriptions automatiques. Ainsi pour chaque nouveau cadre applicatif envisagé, une collecte doit être menée, au moins pour adapter ou entraîner les modèles de langage utilisés par le module de RAP.

Granularité des annotations

Les protocoles d'annotations de corpus peuvent être catégorisés par rapport à deux dimensions : leur proximité de la tâche et leur granularité.

Le degré de proximité permet de balancer d'une part l'utilité immédiate d'une annotation pour construire un système effectif, et d'autre part sa généricité pour pouvoir être appliquée à une autre tâche.

La granularité des annotations permet d'envisager plusieurs niveaux d'évaluation comme nous le verrons dans la section suivante. De plus, en découpant le problème de la représentation du sens en sous problèmes, il est plus facile de gérer le problème de la proximité à la tâche, certains niveaux pouvant être assez génériques alors que d'autres restent directement liés à celle-ci.

C'est ce paradigme d'annotation qui a été choisi dans le projet Européen LUNA [5] servant de cadre à plusieurs de ces travaux. Le paradigme LUNA, décrit dans (Raymond et al., 2007), contient les niveaux d'annotations suivants, certains obligatoires, d'autres facultatifs :

1. segmentation des messages et annotation en mots (obligatoire) ;
2. étiquetage des mots en catégories morpho-syntaxiques et découpage des messages en syntagmes minimaux (obligatoire) ;
3. annotation « à plat » en *concepts* (génériques tels que les entités nommées ou spécifiques à la tâche) sous la forme *attribut-valeur* (obligatoire) ;
4. annotation structurée en prédicat suivant le formalisme *FrameNet* (Baker et al., 1998) (facultatif) ;
5. annotation des coréférences, regroupant à la fois les relations anaphoriques, mais également les relations référentielles et les *bridging relations* telles que définies dans (Vieira et Poesio, 2000) (facultatif) ;
6. annotation en actes de dialogues à partir d'un sous-ensemble de la liste définie dans le projet DAMSL (Core et Allen, 1997) (facultatif).

[5] http ://www.ist-luna.eu

Le caractère *obligatoire* et *facultatif* des différents niveaux sert à définir le niveau minimal d'annotations nécessaire à la construction d'un système automatique. En effet pour la plupart des applications de faible complexité mises en service actuellement, les niveaux 1 à 3 sont suffisants. Par contre pour permettre des capacités de compréhension plus élaborées, telles que celles nécessaires dans les systèmes de dialogue permettant la négociation entre un utilisateur et le serveur de dialogue, une représentation structurée prenant en compte le contexte du dialogue via les actes de dialogue et la gestion des références est nécessaire.

2.4 Comment évaluer un système de compréhension de la parole ?

Le problème de l'évaluation est crucial dans cette étude dans la mesure où le choix entre plusieurs modèles ne peut se faire que sur une évaluation de leurs performances respectives face à une tâche donnée, plutôt que sur leur capacité de description d'un phénomène. L'évaluation a deux principaux objectifs : d'une part optimiser les paramètres des modèles de compréhension sur un corpus donné ; d'autre part comparer entre elles deux approches.

Avant de présenter différentes méthodes d'optimisation et de comparaison, nous allons définir quelles sont les mesures possibles à notre disposition pour évaluer la performance d'un module de compréhension.

2.4.1 Mesures d'évaluation

Les mesures d'évaluation dépendent fortement du type de représentation choisie pour l'annotation sémantique. On peut distinguer deux types d'évaluation : les évaluations de *séquences* de symboles et les évaluations globales à un message.

Evaluation de séquences

Le premier type d'évaluation correspond à une annotation séquentielle calquée sur la séquence de mots prononcée dans le message. Chaque mot ou groupe de mots est associé à un symbole représentant une structure sémantique donnée (paire attribut/valeur, structure prédicative). L'évaluation consiste à aligner dynamiquement la séquence de symboles référence avec celle obtenue automatiquement, puis à estimer le nombre d'insertions, de substitutions et suppressions de symboles nécessaires à cet alignement.

Prenons comme exemple un message dont l'annotation de référence est une séquence de N tokens et dont l'annotation automatique produit une séquence de N' tokens. Si l'alignement entre ces deux séquences a produit i insertions, s substitutions et d suppressions, pour évaluer les performances de l'annotation automatique on peut utiliser les mesures suivantes :

le taux de tokens corrects :

$$T_{cor} = \frac{(N' - i - s) \times 100}{N}$$

les trois taux d'erreurs :

$$T_{ins} = \frac{i \times 100}{N} \quad T_{sub} = \frac{s \times 100}{N} \quad T_{sup} = \frac{d \times 100}{N}$$

le taux d'erreurs global (*token error rate*, *TER*) ou bien l'évaluation globale de la qualité de la séquence (*token accuracy*, *acc*) :

$$T_{TER} = T_{ins} + T_{sub} T_{sup} \quad T_{acc} = 100 - T_{TER} \tag{2.1}$$

Ce type d'évaluation est celui traditionnellement utilisé pour évaluer les systèmes de RAP, les tokens représentant les mots du lexique de reconnaissance. Pour l'évaluation de la compréhension les tokens dépendent de la représentation sémantique choisie.

Evaluation globale

Pour l'évaluation globale on considère que le module de compréhension associe à un message oral une structure sémantique *globale* au message. Il n'y a plus ici de notion séquentielle. Cette structure sémantique peut être composée d'une seule étiquette, comme dans le cas des applications de routage d'appels, cette étiquette représentant la destination du message. Elle peut aussi représenter un ensemble de paires attribut/valeur ou encore des structures prédicatives complexes comme les cadres sémantiques définis dans le formalisme *FrameNet* ou chaque *Frame* est décrite par son type et les *FrameElement* la composant.

On se place ici dans un cadre de classification (ou décision) et non dans un cadre de transcription. Pour évaluer ce processus de décision, on évalue le nombre de *décisions correctes* prises et on le compare au nombre de *choix* possibles présents dans le corpus de référence. Les termes de *décision* et *choix* s'apparentent ici à la production d'un symbole représentant soit la structure sémantique complète associée au message, soit un composant de cette structure. Pour chaque *choix* du corpus référence le système de compréhension peut produire deux résultats : prise de décision ou aucune décision. Chaque décision prise par le système automatique peut recevoir aussi deux étiquettes : décision correcte ou incorrecte.

Par exemple, pour un message dont l'annotation de référence contient l'ensemble de symboles (A, D, E, F), si l'annotation automatique a produit l'ensemble A, B, F, nous aurons ici deux décisions correctes (sur A et F), une décision incorrecte (sur B) et deux absences de décision (sur D et E).

Pour ce type d'évaluation on utilise les mesures de *précision* et *rappel* définies de la

manière suivante :

$$Précision = \frac{nombre\ de\ décisions\ correctes \times 100}{nombres\ de\ décisions\ produites}$$
$$Rappel = \frac{nombre\ de\ décisions\ correctes \times 100}{nombre\ de\ choix\ dans\ la\ référence}$$

Les cas extrêmes de ces mesures sont :
- si un processus de décision produit pour chaque message traité l'ensemble des symboles disponibles dans le système d'annotation sémantique utilisé, le rappel aura une valeur de 100% et une précision très basse, proche de 0 ;
- à l'inverse, si un processus de décision est très strict et ne produit un symbole que lorsqu'il n'y a aucune ambiguïté dans l'interprétation des messages, la précision sera excellente, proche de 100%, mais le rappel aura une valeur très basse.

Ces deux mesures étant antagonistes (maximiser le rappel dégrade la précision et inversement), il est utile de disposer d'une valeur unique pour évaluer un système. On les combine habituellement par le biais de la mesure F (*F-measure*), qui correspond à la moyenne harmonique uniformément pondérée des mesures de précision et de rappel. Elle est définie comme suit :

$$F = \frac{2 * Rappel * Précision}{Rappel + Précision} \tag{2.2}$$

Comme la précision et le rappel, la mesure F prend ses valeurs entre 0 et 100%.

Evaluation Oracle

Quel que soit le type d'évaluation choisie, global ou à base de séquences, on utilise souvent une mesure complémentaire pour évaluer non plus une hypothèse unique mais un ensemble d'hypothèses. Cette mesure s'appelle, dans le cas de l'évaluation d'un taux d'erreur TE, la mesure Oracle pour le taux TE. Elle est définie comme suit : soit une liste L de N hypothèses H_1,H_2,...,H_N, le taux Oracle de la mesure TE pour la liste L, noté $Oracle_{TE}(L)$ est défini par :

$$Oracle_{TE}(L) = \min_{1<=i<=N} TE(H_i) \tag{2.3}$$

Cette valeur permet de mesurer le potentiel de gain atteignable en prenant en compte un ensemble d'hypothèses, en supposant que l'on dispose d'une règle de décision parfaite pour choisir une solution dans cet ensemble, représentée par l'*Oracle*.

2.4.2 Evaluer pour optimiser

L'optimisation d'un système consiste à définir un cycle :

1. réglage des paramètres ;
2. apprentissage des modèles ;
3. évaluation ;
4. analyse manuelle des résultats ;
5. retour vers 1.

Quels que soient les modèles utilisés, un certain nombre de paramètres doivent être définis, le plus souvent de manière totalement empirique. Par exemple pour un modèle d'analyse statistique de type modèle de langage *n-grammes*, l'ordre du modèle, la méthode de repli choisie, le choix du ou des corpus d'apprentissage et de leur éventuel mélange, le degré de tokenisation (*pomme_de_terre* ou *pomme de terre*) sont autant de paramètres à fixer avant l'apprentissage du modèle, à partir du *savoir-faire* du concepteur du système, en l'absence de cadre théorique justifiant un choix précis.

De la même manière, pour un modèle de classification, la représentation des observations (sac de mots ou de *n-grammes*, mesures de confiance données par le module de RAP, contexte de production du message) comme les paramètres de l'algorithme d'apprentissage utilisé (nombre d'itérations, choix du noyau pour les SVM, critères d'arrêt) sont à affiner pour chaque nouveau corpus à traiter.

Dans une perspective d'optimisation le choix de la mesure d'évaluation est très libre. En effet celui-ci doit respecter une seule contrainte : être corrélé avec les performances globales du système mis en place. On peut ainsi se permettre de définir des mesures *ad hoc*, n'ayant aucune interprétation objective, à partir du moment où cette corrélation a été observée sur un corpus représentatif. Par exemple pour l'optimisation d'un système produisant des cadres sémantiques, on peut restreindre l'évaluation, dans un but d'optimisation, à la mesure F sur la détection des éléments composant les frames. Même si cette mesure n'est pas directement interprétable comme une mesure de performance de l'annotation en cadre sémantique (80% des éléments des frames peuvent être corrects, mais peut-être qu'aucun cadre n'est correctement complètement instancié !), elle est évidemment corrélée avec celle-ci et permet d'apporter un niveau de granularité plus fin dans l'évaluation limitant le nombre de paramètres à optimiser en même temps.

2.4.3 Evaluer pour comparer

En ce qui concerne la comparaison entre méthodes, elle doit permettre de déterminer de manière objective les forces et les faiblesses des différentes approches selon le cadre expérimental choisi. Elle doit surtout avoir une certaine généricité afin que les conclusions obtenues puissent être vérifiées sur un autre corpus et dans une certaine mesure dans un autre contexte.

Comme pour les protocoles d'annotation présentés dans le paragraphe 2.1, deux approches sont utilisées dans ce but :

1. évaluation par rapport à la tâche ;
2. évaluation *par niveau*.

Evaluation par rapport à la tâche

Cette évaluation a le principal avantage d'être celle qui aura un impact direct sur les performances d'un système telles qu'elles sont perçues par les utilisateurs. Ainsi, pour une tâche de dialogue, la meilleure évaluation est celle que l'on peut faire à partir d'enquêtes de satisfaction collectées auprès d'utilisateurs du système. Ce type d'évaluation subjective est bien évidemment difficile et coûteux à mettre en place. C'est pour cela que des projets tels que le projet *PARADISE* (Walker et al., 1997) ou *DARPA COMMUNICATOR* (Walker et al., 2001) ont tenté de mettre au point des protocoles d'évaluation objective de systèmes de dialogue, à partir de paramètres extraits automatiquement (nombre de tours de dialogue, de demandes de répétition, de confirmation et négation, *etc*.).

Pour les tâches de fouilles de données nécessitant un module de compréhension (analyse des logs dans les centres d'appel, ou *distillation* d'informations à partir d'émissions diffusées), l'évaluation par rapport à la tâche est plus facile à mettre en œuvre mais pose néanmoins un problème crucial : la nécessité de disposer d'un corpus d'évaluation totalement annoté. Or le but de la fouille de données est de traiter des bases de très grande taille, comme l'audio disponible sur le WEB ou les bases de messages oraux collectés dans les centres d'appel. S'il est possible d'évaluer *a posteriori* la pertinence des résultats retournés, il est très difficile d'en mesurer le rappel par rapport à tout ce qu'il y avait potentiellement à retrouver.

De plus, ce type d'évaluation pose le problème de la *boîte noire* : il y a tellement de paramètres conditionnant les performances d'un système dans un cadre applicatif qu'il est très difficile de tirer des conclusions sur les avantages ou les inconvénients des différents modèles pouvant être utilisés dans les composants du système.

Evaluation par niveau

Dans ce type d'analyse on évalue toutes les strates du processus de compréhension, en commençant par les performances du module de RAP. Ensuite différentes sous-tâches sont identifiées comme étant partie prenante du processus de compréhension. Par exemple la tâche d'extraction d'entités (telles que les entités nommées), la désambiguïsation sémantique des formes verbales, l'identification et la résolution des relations telles que les anaphores pronominales, le suivi d'entité, *etc*. Toutes les tâches de l'analyse syntaxique et sémantique de surface peuvent intervenir dans ces évaluations.

L'avantage de ce type d'évaluation à la granularité fine est de pouvoir comparer les performances de différentes méthodes sur un point précis, relativement indépendant vis-à-vis de la tâche finale, ce qui permet de limiter le nombre de paramètres à envisager sur chaque point et d'augmenter la généricité de l'évaluation. Cependant il est difficile de savoir si les gains obtenus sur l'une des sous-tâches considérées seront significativement utiles pour la résolution de la tâche globale.

2.4.4 Discussion

L'évaluation est cruciale dans l'étude de tout modèle à base d'apprentissage sur corpus, c'est elle qui permet d'entrer dans le cycle apprentissage-évaluation-optimisation des modèles. C'est notamment grâce au développement de campagnes d'évaluation à partir des années 80 (principalement les campagnes DARPA) ainsi qu'à la production de corpus de référence sur lesquels tout le monde peut évaluer son système, que le traitement automatique de la parole a obtenu des succès notables depuis une vingtaine d'années. Ces succès peuvent se constater par le transfert de technologies vocales vers des applications industrielles (dictée vocale, serveurs de dialogue, serrure vocale).

Ce recours systématique à l'évaluation a cependant aussi ses travers, surtout lorsqu'elle concerne des tâches qui ne sont pas toujours aussi clairement définies que la transcription en mots d'un signal de parole. Les principaux problèmes posées par l'évaluation dans le contexte de la compréhension de messages vocaux sont de deux types : problèmes de disponibilité des corpus ; problèmes de la granularité de l'évaluation.

Disponibilité des corpus

Les sections 2.1 et 2.3 ont souligné d'une part les difficultés que comportait l'annotation formelle du *sens* d'un message en dehors d'un cadre applicatif précis et d'autre part les problèmes posés par l'acquisition et l'annotation de corpus. Ces difficultés font que contrairement aux corpus d'évaluation de la tâche de transcription pour laquelle il existe de très nombreux corpus avec de multiples conditions (parole lue, parole conversationnelle, émissions diffusées, *etc.*), il existe très peu de corpus de parole disponible (hors corpus industriel) avec un étiquetage pouvant être utilisé pour étudier le problème de l'extraction du sens.

L'exception notable, pour la langue anglaise, est le corpus ATIS (Air Travel Information Service), dont le projet s'est étendu de 1989 à 1994. Ce corpus de laboratoire, pourtant ancien, est encore largement utilisé actuellement par manque d'alternatives, par exemple dans les études récentes (He et Young, 2003; Raymond et Riccardi, 2007). Cependant il a de nombreuses restrictions : parole non téléphonique, peu spontanée, domaine sémantique restreint, pas de contexte, annotations limitées aux requêtes SQL décrivant les demandes des locuteurs.

En Europe, malgré deux projets consacrés à la compréhension de la parole, le projet Esprit SUNDIAL (Speech Understanding and DIALog) de 1988 à 1993 (Peckham, 1993) et le projet ARISE (Automatic Railway Information Systems for Europe) de 1996 à 1998 (den Os et al., 1999), aucun corpus de référence n'est disponible pour le français, à part le corpus MEDIA, déjà évoqué dans ce document, et sur lequel s'appuie une partie du travail de cette habilitation.

Ce manque de corpus et de standard d'annotation a pour conséquence que l'évaluation la plus communément utilisée pour évaluer un système de traitement de messages oraux est la mesure du taux d'erreurs dans les transcriptions automatiques en mots.

Cette mesure, si elle a le mérite d'être généralement corrélée avec les mesures de performance des systèmes utilisant les transcriptions automatiques, est loin d'être satisfaisante car elle donne le même poids à tous les mots et tous les segments d'un message de parole, qu'ils soient ou non utiles à l'obtention du sens.

C'est pour répondre à cette carence que le projet Européen LUNA a démarré en 2006, un des résultats concrets de ce projet étant la mise à disposition de corpus et d'annotations sémantiques sur plusieurs niveaux dans le cadre d'applications de dialogue pour trois langues européennes : le français, l'italien et le polonais.

Granularité de l'évaluation

Si l'évaluation globale est finalement le but à atteindre avant la mise en service d'une application, l'évaluation par niveau est de loin l'évaluation la plus utilisée dans la communauté scientifique. Cependant le principe de l'annotation de faible granularité a aussi ses inconvénients : à force de découper une compétence linguistique en sous-tâches, on peut se poser la question de la pertinence des sous-tâches définies, surtout si elles sont envisagées comme des buts en soi. Ce point est justement soulevé par Yvon dans (Yvon, 2006) (§1.4.1 pp. 40-41), en prenant comme exemple l'annotation de l'emploi référentiel ou non référentiel du pronom *it* en anglais sans tentative de résolution des co-références. En effet, si l'on peut considérer cette localisation comme une trace du processus global de résolution et à ce titre comme une mesure d'évaluation dans un but d'optimisation, il est douteux de la considérer comme un but en soi, les systèmes gérant le problème de la résolution n'effectuant pas cette décomposition.

Un phénomène similaire a été annoté dans le corpus MEDIA : dans un tour de parole, tous les liens référentiels vers des entités situés dans d'autres tours de parole sont annotés. Une des mesures d'évaluation consiste à détecter ces liens, sans pour autant les résoudre. Pour les besoins de l'évaluation nous avons développé au LIA un système fondé sur les modèles CRF se contentant d'effectuer cette détection, sans résolution, et avons obtenus de bons résultats pour la tâche (Denis et al., 2007). Cependant je ne suis pas sûr que les résultats flatteurs obtenus par notre système ne soient d'aucune utilité pour résoudre le problème autrement plus ardu de la résolution fine de ces références !

Enfin une dernière critique que l'on peut faire à l'évaluation par niveau est l'absence de garantie qu'un gain obtenu sur un des niveaux ait un impact sur les performances globales du système développé. Même si les performances sont corrélées, il est difficile d'évaluer la signifiance d'un gain pour les performances globales.

Chapitre 3

Les corpus

Sommaire

3.1 Les corpus « écologiques »

Comme mentionné dans le paragraphe 2.3, les corpus *écologiques* contiennent des données *réalistes*, c'est à dire provenant de sources *a priori* non destinées à être traitées automatiquement par une machine. Dans cette famille de corpus nous présentons dans ce document les travaux réalisés sur un corpus d'émissions radiophoniques, le corpus ESTER et un corpus provenant d'un centre d'appel et contenant des sondages téléphoniques, corpus accessible grâce à une collaboration avec France Télécom R&D.

3.1.1 Corpus de données radiodiffusées ESTER

Le corpus ESTER a été développé dans le cadre du programme d'évaluation Technolangue EVALDA sur la période 2004-2005. Il contient environ 100 heures d'émissions radiophoniques. Ces émissions sont transcrites orthographiquement et incluent les tours de parole avec l'identité des locuteurs, les conditions acoustiques et divers évènements

tels que la présence de bruits. Ces annotations sont définies par le logiciel Transcriber (Barras et al., 2001). Le corpus est composé de 5 radios différentes : Radio France International (RFI), France Inter, France Info, Radio Classique et la Radio Télévision Marocaine (RTM, en langue française). Il se décompose en 82 heures d'apprentissage, 8 heures de développement et 10 heures de test.

La tâche principale évaluée dans ESTER est la transcription en mots. Cependant, afin de ne pas se limiter au seul taux d'erreurs sur les mots, une tâche d'extraction d'*entité nommée* a également été prévue sur ce corpus. Avant de présenter en détail le jeu d'entité nommée utilisé dans ESTER, le paragraphe suivant présente les problématiques liées à cette tâche.

Entités Nommées

Les *entités nommées* ont été popularisées par les conférences sur l'extraction de contenu (*MUC - Message Understanding Conference*). La règle générale permettant de décider si une expression est ou n'est pas une entité nommée se fonde sur le caractère d'unicité du concept représenté par l'expression. Les inévitables ambiguïtés produites par une définition aussi générale sont traitées par un ensemble de recommandations à l'usage des annotateurs de corpus (Chinchor et Robinson, 1998).

Ces entités nommées peuvent représenter des noms de personnes, de lieu ou d'organisation, ainsi que des expressions temporelles (dates et temps) ou des quantités numériques (valeurs monétaires, pourcentages, *etc.*). Les entités nommées contenant des noms propres peuvent prendre des formes très diverses : soit composées de noms propres isolés (*UNESCO*), de noms propres associés à des noms communs modifiant leur sens (*Place Vendôme*) ou, encore, de groupes nominaux contenant uniquement des noms communs (*Banque Populaire*). Ces entités nommées vont poser deux types de problèmes aux systèmes devant les détecter à partir d'un flux audio :

1. pour celles composées uniquement de noms propres (patronymes, toponymes, etc.) le problème principal sera celui de leur recensement, de leur caractérisation et de la construction des formes phonétiques correspondantes ;
2. pour celles qui sont composées (en tout ou partie) de mots du vocabulaire commun, il s'agira de détecter leur appartenance à une entité nommée, ne serait-ce que pour produire une capitalisation correcte de leur graphie en sortie du système de reconnaissance de parole.

La tâche de détection des entités nommées est présente dans de nombreux programmes d'évaluation tels que les campagnes d'évaluation *Message Understanding Conference* MUC-6 et MUC-7 ; les évaluations multilingues (*Multilingual Entity Task*) MET-1 et MET-2, les campagnes DARPA sur la transcription de données audiovisuelles (*DARPA Broadcast News HUB-5*) ou encore les *Conferences on Natural Language Learning* (CoNLL).

La tâche d'extraction d'entités nommées consiste à localiser dans un texte puis à identifier (par exemple avec un balisage de type SGML) des expressions ayant été au préalable étiquetées comme nom de personne, de lieu ou d'organisation.

Cette tâche rentre dans le cadre de la problématique de la compréhension de la parole telle qu'elle est définie dans le chapitre précédent : des informations conceptuelles sont associées à des séquences de mots, chaque concept correspond à une entité nommée représentée par une paire attribut/valeur.

Les entités nommées dans ESTER

Le jeu d'étiquettes représentant les entités nommées dans le corpus ESTER contient 30 catégories groupées en 8 macro-classes :

- personnes (**pers**) : humain, personnage de fiction, animal ;
- lieux (**lieu**) : géographique, adresse (électronique et postale), numéros de téléphone ;
- organisations (**org**) : politique, commerciale, associative ;
- groupe geo-socio-politique (**gsp**) : clan, famille, nation, région administrative ;
- montants (**montant**) : durées, argent, dimensions, température, âge, poids et vitesse ;
- temps (**temps**) : expression relative et absolue, heure ;
- produits (**prod**) : art, imprimé, prix et véhicules ;
- bâtiments (**bat**) : immeubles, monuments.

Les données sont divisées en trois sous-ensembles : apprentissage (84%), développement (8%) et corpus de test (8%). Il y a une différence de six mois entre le corpus d'apprentissage et le corpus de test alors que le corpus de développement est identique, d'un point de vue temporel, à celui du test. Il y a aussi deux stations de radio présentes dans le test qui n'étaient pas dans le corpus d'apprentissage.

La principale caractéristique du corpus ESTER d'entités nommées est la taille du jeu d'étiquettes et la forte ambiguïté dans la définition des catégories d'entités (par exemple régions administratives et lieux géographiques) : 83% des entités du corpus de développement se trouvent dans le corpus d'apprentissage et 40% d'entre-elles sont ambiguës ; 61% des entités du corpus de test sont dans le corpus d'apprentissage dont 32% sont ambiguës.

Un exemple extrait du corpus ESTER avec les marqueurs d'entités nommées est donné dans la table 3.1.

3.1.2 Corpus d'enquêtes d'opinions France Télécom

Extraire d'un texte des expressions subjectives représentant l'opinion de l'auteur sur un sujet précis a récemment fait l'objet d'un grande attention de la part de la communauté du traitement automatique du langage, comme le montre l'atelier d'ACL 2006 *Sentiment and Subjectivity in Text* ou encore la campagne d'évaluation DEFT'07 portant sur l'analyse de critiques de livres, films ou spectacles ainsi que sur des commentaires d'hommes politiques.

<entity type="org.com"> *France Inter* **</entity>** il est **<entity type="time.hour">** *7 heures 4* **</entity>** **<entity type="org.com">** *France Inter* **</entity>** a décidé pendant **<entity type="amount.phy.dur">** *3 semaines* **</entity>** d'ouvrir son antenne à **<entity type="org.non-profit">** *la chaîne de l'espoir* **</entity>** **<entity type="org.non-profit">** *la chaîne de l'espoir* **</entity>** c'est une association créée il y a **<entity type="time.date.rel">** *10 ans* **</entity>** qui se mobilise pour aider les enfants du monde entier qui sont condamnés à mourir ou à rester infirmes toute leur vie parce que les soins coûtent chers ou parce qu'ils ne peuvent pas être soignés dans leur pays l'association a été créée par un professeur en chirurgie de l' **<entity type="fac">** *hôpital Broussais* **</entity>** à **<entity type="gsp.loc">** *Paris* **</entity>** **<entity type="pers.hum">** *Alain Deloche* **</entity>** et grâce à l'aide de personnalités du monde médical et à différents partenaires publics ou privés , **<entity type="org.non-profit">** *la chaîne de l'espoir* **</entity>** a pu soigner en **<entity type="gsp.loc">** *France* **</entity>** 1300 enfants venus d'une trentaine de pays ils sont hébergés chez des bénévoles ils sont opérés , soignés , et ensuite raccompagnés chez eux , quand ils sont guéris d'autres actions de solidarité se déroulent directement dans les pays où se trouvent les enfants ; en tout , depuis **<entity type="time.date.abs">** *1995* **</entity>** cette association a permis de sauver 5000 enfants

TAB. *3.1:* *Extrait du corpus ESTER avec les marqueurs d'entités nommées*

Ce domaine a donné lieu à de nombreuses publications (Wiebe et al., 2005; Choi et al., 2005) portant principalement sur deux aspects : la détection automatique d'opinions à partir d'avis rédigés par des consommateurs (Popescu et Etzioni, 2005) et d'autre part l'analyse de la subjectivité d'une phrase pour les systèmes de résumé automatique ou de question/réponse (Riloff et Wiebe, 2003).

Le corpus d'enquêtes de France Télécom décrit dans ce paragraphe concerne le premier cadre applicatif : la détection automatique d'opinions à partir de sondage d'utilisateurs. Il s'agit ici de sondages téléphoniques effectués par France Télécom auprès d'utilisateurs réels. Une des principales caractéristiques de cette étude est la détection d'opinions à partir de messages vocaux, contenant de la parole complètement spontanée, collectée dans des conditions réelles.

Description du corpus

Les personnes sont invitées par un court message à appeler un numéro gratuit qui leur permet d'exprimer leur satisfaction vis à vis du service client qu'ils ont récemment appelé. En composant ce numéro, le message vocal suivant les invite à laisser un message :
«*[. . .] Vous avez récemment contacté notre service clientèle. Nous souhaitons nous assurer que vous avez été satisfait de l'accueil et de la suite donnée à votre appel. Vous pourrez me laisser votre réponse après le top sonore. N'hésitez pas à me faire part de tous vos commentaires et de vos suggestions sur notre service, ceux-ci nous aideront à nous améliorer. Nous vous remercions*

de votre aide et nous restons à votre disposition. Laissez votre message après le signal sonore. »

Du fait que les messages ont été enregistrés à l'origine dans l'optique d'un traitement par opérateur, aucune consigne de nature à faciliter le traitement automatique n'a été donnée : pas de conseils sur le mode d'élocution, question ouverte et même incitation à laisser des commentaires. Ainsi, les messages recueillis sont *réalistes* et de longueur variable (d'une dizaine à plusieurs centaines de mots). Pour cette étude un ensemble de 1779 messages, collectés sur une période de 3 mois, a été transcrit manuellement au niveau mots, opinions et marqueurs (indications de disfluences et marqueurs discursifs).

Le lexique de l'application comporte environ 4500 mots .
Ce corpus a été divisé en trois sous-corpus. Environ 50% des phrases composent le corpus d'apprentissage, 33% le corpus de développement et 17% le corpus de test.

Étiquettes sémantiques

L'analyse de la satisfaction des utilisateurs par l'équipe d'analyse des sondages se fait selon trois dimensions : la qualité de l'accueil (notée *accueil*), la rapidité d'accès au service (notée *attente*) et enfin l'efficacité du service (notée *efficacité*). Cette dernière dimension est la plus représentée dans le corpus, elle concerne à la fois l'évaluation des réponses aux attentes des utilisateurs (est ce que le problème a été réglé ?) mais aussi la qualité des informations données. Chaque expression subjective peut recevoir deux polarités : *positive* et *négative*. Nous avons donc un total de 6 étiquettes pour caractériser les expressions subjectives du corpus.

Dans la transcription manuelle, au sein de chaque message, ces expressions sont indiquées par des balises. Nous disposons d'un corpus de segments, chacun porteur d'une ou plusieurs opinions particulières. Le but du module de compréhension automatique est de retrouver ces segments et de les étiqueter avec l'une des 6 étiquettes.

Un exemple de message avec les balises de référence est donné dans le tableau 3.2.

« *oui c'est monsieur NOMS PRENOMS j'avais appelé donc le service client ouais* **<seg label=accueil,pos>** *j'ai été très bien accueilli* **</seg>** *des* **<seg label=efficacité,pos>** *bons renseignements* **</seg>** *sauf que* **<seg label=efficacité,neg>** *ça ne fonctionne toujours pas* **</seg>** *donc je sais pas si j'ai fait une mauvaise manipulation ou y a un problème enfin voilà sinon* **<seg label=efficacité,pos label=accueil,pos>** *l'accueil était et les conseils très judicieux* **</seg>** *même si* **<seg label=efficacité,neg>** *le résultat n'est pas n'est pas là* **</seg>** *merci au revoir* »

TAB. 3.2: *Exemple de message du corpus d'opinions France Télécom contenant plusieurs opinions avec les marqueurs de segments*

Nb concept par message	Répartition (% corpus)	Taille moyenne (nb mots)
0	19.2	61.0
1	51.3	40.3
2 et plus	29.5	60.8

TAB. 3.3: *Répartition des messages dans le corpus d'opinions France Télécom en fonction du nombre de concepts exprimés*

Caractéristiques des messages

Le tableau 3.3 montre que la taille moyenne d'un message n'augmente pas en fonction du nombre de concepts exprimés. Il faut même autant de mots, si ce n'est plus pour n'exprimer aucun des concepts recherchés que pour en exprimer un. Cela s'explique, d'une part, par le *hors-sujet* du locuteur qui peut s'exprimer autant sur la cause de son problème ou sur sa situation personnelle que sur son ressenti vis à vis du service client et d'autre part, par le fait qu'un même segment de message peut être support de plusieurs critères.

Concernant le nombre moyen de mots nécessaires pour exprimer un concept, les concepts évoquant une polarité négative nécessitent souvent plus de mots que les concepts évoquant une polarité positive.

Ainsi : « *très bon accueil super très bien merci* » sera étiquetée *satSerAcc* pour *satisfait du service concernant l'accueil*, alors que « *j' ai demandé à supprimer l' option actualité or j' ai reçu depuis encore des hum des appels je vous remercie de les annuler* » sera étiquetée *insatSerEff* pour *insatisfait du service concernant l'efficacité*.

Un des problèmes que posent ces messages est qu'un même concept peut être vu plusieurs fois dans un message avec des opinions contraires. Cela se rencontre quand la personne n'est pas entièrement satisfaite (*e.g.* : satisfaite du service client mais pas du résultat) ou qu'une notion temporelle rentre en jeu dans son discours. Un exemple de ce type de message est donné dans le tableau 3.2.

3.2 Les corpus « de laboratoire »

Les corpus de laboratoire sont des corpus collectés auprès de sujets sachant qu'ils participent à une expérience. Les locuteurs reçoivent un scénario à suivre, plus ou moins explicite, et « jouent » leurs rôles face à un système automatique ou bien un système simulé de type *Magicien d'Oz*. Ces corpus sont utiles lors du développement d'une application ainsi que pour évaluer un système existant.

3.2.1 Le corpus France Télécom PLANRESTO

PLANRESTO est une application de dialogue développée à France Télécom R&D à Lannion implémentant le moteur d'interprétation et de dialogue Artimis (Sadek et al., 1997). Cette application permet la recherche d'information et la réservation de restaurants sur Paris. Dans le cadre de l'amélioration et de l'évaluation de cette application un corpus a été collecté par France Télécom R&D auprès d'utilisateurs volontaires. Cette collecte s'est faite en utilisant une version opérationnelle de PLANRESTO, il ne s'agit donc pas d'une simulation par Magicien d'Oz.

« je cherche un restaurant euh je cherche un restaurant au métro du côté du métro La Tour Maubourg s'il-vous-plaît »

TAB. *3.4:* *Exemples de message du corpus PLANRESTO de France Télécom*

A la suite de cette collecte le corpus a été divisé en trois ensembles : un corpus d'apprentissage comprenant 13000 messages, un corpus de développement contenant 4000 messages et un corpus de test contenant 1500 messages. Chacun de ces corpus a été transcrit manuellement en mots et étiqueté en concepts.

Le corpus PLANRESTO utilisé dans cette étude utilise un ensemble de 59 concepts tels que : marqueurs d'actes communicatifs, marqueurs linguistiques, classes spécifiques, etc. La liste complète est donnée dans le tableau 3.5.

Ces concepts sont utilisés en entrée du système ARTIMIS fondé sur le formalisme logique KLONE (Brachman et Schmolze, 1985). Ils représentent les unités sémantiques élémentaires qui, extraites à partir du texte, permettent la construction de l'interprétation sémantique puis celle de l'interprétation contextuelle. Certains concepts sont reliés à la gestion du dialogue (ex : confirmation ou contestation) et d'autres au domaine d'application (ex : lieu ou date). Dans l'application PLANRESTO, un concept est représenté par une paire attribut/valeur.

Ces 59 concepts sont groupés en un ensemble de 28 classes, chaque classe correspondant à un groupe de concepts partageant des propriétés sémantiques communes.

3.2.2 Le corpus MEDIA

La plupart des applications de dialogue homme-machine mises en service actuellement peuvent être vues comme une interface entre un utilisateur et une base de données. Le but du dialogue est de remplir tous les champs d'une requête qui va être adressée à la base de données. Dans ce cadre les concepts sémantiques sont de 3 types : les concepts relatifs au type de la requête ; les concepts relatifs aux valeurs qui instancient les paramètres de la requête ; et enfin les concepts relatifs à la conduite du dialogue. La campagne d'évaluation MEDIA (Bonneau-Maynard et al., 2005) (programme Technolangue/Evalda) se place dans ce cadre applicatif à travers la simulation d'un système d'accès à des informations touristiques et des réservations d'hôtel. Un corpus de 1250

Concepts avec valeurs

Lieux	un lieu
Prix	un prix
Specialite	une spécialité culinaire
valeur(ord)	un ordinal
valeur(card)	un cardinal

Classe Spécifiques

claAdresse	
claAmbiance	
claArrondissement	
claCapaciteAccueil	
claConnexion	
claEspacesVerts	
claHautsLieuxReligieux	
claHoraire	
claInformation	
claLieu	
claMessage	
claMusees	
claNom	
claPlaces	
claPrix	
claPrixExterne	
claQuartiers	
claRestaurant	
claSpecialite	
claStations	
claTel	

verbe être

vb(neg_rmoi)	
vb(pos_rmoi)	

Aucun concept (hors focus)

BCK	Aucun concept

marqueurs d'actes communicatifs

ma(aide)	demande d'aide
ma(end_of_session)	demande à quitter
ma(raz)	demande de remize à zéro
ma(reeng_Diag)	
ma(repeter)	demande de répétition
ma(modeGuide)	demande à être guidé
ma(petiteRelance)	

marqueurs linguistiques

ml(contest)	contestation
ml(inver_v_suj)	inversion verbe/sujet
ml(neg_pre)	négation pécédent un verbe
ml(non)	réponse négative
ml(object)	pronom à la troisième personne
ml(ord(prec))	ordinal indiquant la précédence
ml(ord(svt))	ordinal indiquant le suivant
ml(ord(dernier))	ordinal indiquant le dernier
ml(tous)	toutes les réponses
ml(oui)	réponse positive

opérateurs modaux

op(neg_krif_auditeur)	
op(neg_krif_locuteur)	
op(pos_kif_auditeur)	
op(pos_kif_locuteur)	
op(pos_krif_auditeur)	
op(pos_krif_locuteur)	

Divers

consulter	
dans	
peu_importe	
retour	
mini	
maxi	
utilisateur_regulier	

TAB. 3.5: *Liste des concepts de l'application PLANRESTO*

dialogues a été enregistré par ELDA selon un protocole de *Magicien d'Oz* : 250 locuteurs ont effectué chacun 5 scénarios de réservation d'hôtel avec un système de dialogue simulé par un opérateur humain.

Un exemple de dialogue est donné dans la table 3.6.

Ce corpus a ensuite été transcrit manuellement, puis annoté sémantiquement selon un dictionnaire sémantique de concepts mis au point par les partenaires du projet MEDIA (Bonneau-Maynard et al., 2005).

Représentation sémantique

Le dictionnaire sémantique utilisé pour annoter le corpus MEDIA (Bonneau-Maynard et al., 2005) permet d'associer 3 types d'information à un mot ou un groupe de mots :

– tout d'abord une paire attribut-valeur, correspondant à une représentation sémantique *à plat* d'un énoncé ;

woz> *bienvenue sur le serveur MEDIA système d' informations touristiques et de réservation d' hôtel quelle information souhaitez-vous*
spk> *oui j' aimerais réserver trois nuits à Marseille du quinze au dix-huit mars s' il vous plaît*
woz> *vous souhaitez faire une réservation à Versailles du quinze du quinze au dix-huit mars*
spk> *non non non à Marseille Marseille pas Versailles Marseille Marseille*
woz> *vous souhaitez faire une réservation à Marseille du quinze au dix-huit mars*
spk> *oui*

TAB. 3.6: *Exemple de dialogue extrait du corpus MEDIA*

– puis un spécifieur qui permet de définir des relations entre les attributs et qui par conséquent peut être utilisé pour construire une représentation hiérarchique de l'interprétation d'un énoncé ;
– enfin une information sur le *mode* attaché à un concept (positif, affirmatif, interrogatif ou optionnel).

n	W^{c_n}	c_n	*mode*	*spécifieur*	*valeur*
1	*je vais réserver*	command-tache	+		reservation
2	*dans cet hôtel hôtel Richard Lenoir*	nom-hotel	+		richard lenoir
3	*six*	nombre-chambre	+	reservation	6
4	*chambres individuelles*	chambre-type	+		simple
5	*pour le trente et un mai*	temps-date	+	reservation	31/05
6	*deux jours et deux nuits*	sejour-nbNuit	+	reservation	2

TAB. 3.7: *Exemple d'annotation MEDIA sur le message : « bon ben écoutez je vais réserver dans cet hôtel hôtel Richard Lenoir donc six chambres individuelles pour le trente et un mai deux jours et deux nuits hein »*

La table 3.7 présente un exemple de message annoté du corpus MEDIA. La première colonne correspond au numéro du segment dans le message, la deuxième colonne à la chaîne de mots W^{c_n} porteuse du concept c_n contenu dans la troisième colonne. Les colonnes 4, 5 et 6 contiennent le mode, le spécifieur et la valeur du concept c_n dans la chaîne W^{c_n}. Le dictionnaire sémantique MEDIA contient 83 attributs, auxquels peuvent s'ajouter 19 spécifieurs de relations entre attributs. Les attributs sont dérivés de la base de données associée à l'application MEDIA.

Les spécifieurs sont dépendants de l'attribut et permettent de modifier ou de préciser la signification associée à l'attribut suivant le cas. Par exemple, le spécifieur *reservation* associé à l'attribut *paiement-montant* permet de préciser que la chaîne W^{c_n} associée se rapporte au montant de la réservation en cours.

Enfin, des valeurs normalisées ont été adjointes au couple attribut-spécifieur. Elles ont été définies dans un dictionnaire sémantique avec trois configurations possibles :
– une liste de valeurs (par exemple *singulier*, *pluriel*, *etc.*) ;
– des expressions régulières (pour les dates par exemple) ;
– des valeurs ouvertes (pour les entités nommées principalement).

Le corpus collecté a été découpé en plusieurs lots. Nous utilisons dans cette étude les 4 premiers lots comme corpus d'apprentissage, soit 720 dialogues contenant environ

12 000 messages d'utilisateurs et le lot 5 comme corpus de tests contenant 200 dialogues avec 3000 messages d'utilisateurs.

Annotation avec le contexte du dialogue

La campagne d'évaluation MEDIA a été scindée en deux parties, chacune ayant donnée lieu à une annotation spécifique du corpus : la première, appelée *évaluation hors-contexte*, considère chaque tour de parole comme indépendant, aucun contexte de dialogue n'est pris en compte. Pour produire cette annotation les annotateurs humains recevaient les tours de parole dans un ordre aléatoire, ils ne devaient tenir compte d'aucun historique dans leur processus d'annotation.

La deuxième annotation, appelée *évaluation en-contexte*, a été faite à partir de l'annotation hors-contexte en rajoutant cette fois le contexte du dialogue. Les dialogues étaient ainsi présentés aux annotateurs dans leur globalité.

Le contexte du dialogue influe sur trois paramètres :

1. les attributs ; par exemple si un tour de parole ne contient que l'énoncé « quatre », l'annotation hors-contexte se contentera de préciser qu'il s'agit d'une entité numérique alors que l'ajout du contexte permettra de spécifier que cette entité représente un nombre de chambres ou un nombre de nuits ;
2. les spécifieurs ; ils sont une trace de la structure prédicative représentant le sens d'un énoncé, lorsque ce sens est exprimé sur plusieurs tours de parole, les spécifieurs sont hérités des tours précédents ;
3. les liens référentiels.

Concernant la référence, l'annotation hors contexte s'est limitée à annoter la présence d'une expression référentielle grâce à un trait d'attribut *lienRef* raffiné par la catégorie de l'expression. Les différents raffinements retenus (appelés *spécifieurs*) sont proches des catégories définies dans le projet *Reference Annotation Framework*, RAF (Salmon-Alt et Romary, 2004) (voir tableau 3.2.2). Lors de l'évaluation en contexte les liens vers les entités concernées par des *lienRef* sont annotés.

Spécifieur	Signification	Expressions référentielles
coRef	coréférence : l'expression référentielle désigne son référent par référence directe	pronoms, définis, démonstratifs
elsEns	élément-ensemble : l'expression référentielle désigne son référent en vertu de propriétés sémantiques ou indexicales qui l'opposent à d'autres entités dans un ensemble	ordinaux, superlatifs, relatives, certains pronoms démonstratifs
coDom	co-domaine : l'expression référentielle désigne son référent grâce à un marqueur linguistique d'altérité	altérités

TAB. *3.8:* *Types de spécifieurs MEDIA pour les expressions référentielles*

Afin de réduire le coût d'annotation, seules les expressions référentielles dont la résolution dépasse le cadre de l'énoncé ont été annotées. Cela exclut dès lors les référents

dont l'antécédent a été introduit dans le même énoncé, les entités nommées et les indéfinis. En revanche, les articles définis ont été annotés systématiquement, du moins pour les entités relevant de la tâche.

n	W^{c_n}	c_n	*mode*	*spécifieur*	*valeur*
1	ils	lienRef	+	coRef	pluriel
2	proches d'	loc-distanceRelative	?		proche
3	un parc	loc-lieuRelatif	?	general	parcJardin

TAB. *3.9:* *Exemple d'annotation hors contexte MEDIA sur le message : « et j' aimerais savoir s' ils sont proches d' un parc »*

n	W^{c_n}	c_n	référence	*mode*	*spécifieur*	*valeur*
1	ils	lienRef	guillermo champ-mars pullman	+	coRef	pluriel
2	proches d'	loc-distanceRelative		?	hotel	proche
3	un parc	loc-lieuRelatif		?	general-hotel	parcJardin

TAB. *3.10:* *Exemple d'annotation en contexte MEDIA sur le message : « et j' aimerais savoir s' ils sont proches d' un parc » énoncé après le prompt : « je vous propose trois hôtels hôtel Guillermo hôtel du champ de mars hôtel Pullman »*

L'accord inter-annotateur

Afin de vérifier la qualité du corpus, une évaluation a été faite consistant à mesurer et valider l'accord entre les différents annotateurs. La mesure utilisée permettant de valider cet accord (appelé *IAG*) est la mesure *Kappa* k telle que (Carletta, 1996) :

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{3.1}$$

avec $P(A)$ correspondant au nombre de fois où les annotateurs sont d'accord par rapport au nombre d'annotation, et $P(E)$ la probabilité que les annotateurs aient effectué la bonne annotation par chance. Ici $P(E)$ est égale à $\frac{1}{145}$. Le nombre 145 correspondant au nombre de paires *attribut-spécifieur* observées sur l'ensemble du corpus. Il est communément admis dans la littérature qu'une mesure *Kappa* supérieure à 0.8 (soit 80% ici) est considérée comme bonne.

Les mesures d'IAG effectuées lors de la mise au point du corpus MEDIA sont d'environ 90% (Bonneau-Maynard et al., 2005). Cette évaluation tend à montrer que les annotations sont assez homogènes, et donc à valider la qualité des annotations dans le corpus.

Nous pouvons voir dans le tableau 3.11 les résultats des différentes IAG. Ces résultats sont satisfaisant à partir de la quatrième mesure d'IAG, celle-ci correspondant

$N°$ IAG	1	2	3	4	5	6	7
Nb Dialogues	20	8	10	10	10	10	10
Nb tours util.	310	125	183	165	137	106	163
Nb segments	667	459	478	372	455	342	459
Accord (%)	66.1	76.2	78.4	89.5	83.1	83.9	87.8

TAB. *3.11:* *Exemple de message annoté en-contexte du corpus MEDIA*

au moment où les annotatrices ont eu fini d'annoter les 3 premiers corpus d'apprentissage. Les dernières IAG ont été faites sur les corpus de développement et de test, afin de valider les résultats obtenus par les participants lors de la campagne d'évaluation.

3.3 Les corpus provenant d'applications mises en service

Depuis le déploiement en 2000, sur une très grande échelle, du système de dialogue automatique de l'opérateur AT&T *How May I Help You ?* (HMIHY) (Gorin et al., 1997), plusieurs autres systèmes acceptant également la parole spontanée ont été mis en service, notamment en France avec les applications *10 14* et *3000* de France Télécom. Bien qu'il reste de nombreux problèmes scientifiques non résolus (tels que des problèmes de robustesse par exemple), la conception et le déploiement de tels systèmes suivent maintenant un cycle industriel.

Ces corpus ont deux avantages majeurs : d'une part leur quantité est quasiment illimitée, tant que l'application est en service il suffit de collecter les traces des dialogues ; d'autre part ils contiennent de *vrais* dialogues, non simulés, avec des utilisateurs novices ou expérimentés, et la parole est très spontanée. Du point de vue de la recherche académique, ces corpus sont une formidable occasion de pouvoir observer de nombreux exemples de phénomènes encore mal modélisés en traitement automatique de la parole tels que les disfluences, la robustesse aux bruits et aux accents et enfin les énoncés inattendus ou hors domaine.

Les inconvénients liés à ces corpus résident dans leur faible complexité dialogique, due aux contraintes de robustesse des systèmes déployés à une grande échelle et dans l'annotation sémantique *ad hoc* des messages, très proche des fonctionnalités de l'application.

Enfin ces corpus sont difficilement accessibles, car liés par des contraintes commerciales de protection de la vie privée et de droit industriel. Grâce à des partenariats de recherche avec deux grands opérateurs de télécommunication, AT&T aux Etats-Unis et France Télécom en France, nous avons eu la possibilité de travailler sur deux corpus de très grande taille provenant de deux applications de gestion de clientèle : pour l'anglais le corpus *How May I Help You ?*^{tm}, pour le français le corpus *FT3000*.

3.3.1 Le corpus d'AT&T *How May I Help You ?*

Le corpus utilisé dans cette étude contient 130 000 messages collectés auprès de vrais utilisateurs du service *How May I Help You ?* (HMIHY) de AT&T. HMIHY (Gorin et al., 1997) est une application de relation clientèle gérant les requêtes et les plaintes des clients d'AT&T à propos de leurs factures téléphoniques. Ce système repose sur une stratégie de dialogue à initiative mixte : au tout début du dialogue le système demande à l'utilisateur d'exprimer sa requête de manière libre par le simple message : « *How may I help you ?* »[1]. A la suite de sa réponse un dialogue, généralement assez court, permet à l'utilisateur de s'identifier et de préciser sa requête. Finalement l'utilisateur est redirigé vers un sous-dialogue automatique si la requête est traitable par la machine ou vers un opérateur humain si la requête n'est pas couverte par le système.

L'annotation sémantique disponible dans le corpus HMIHY est une annotation en *types d'appel*. Le principe de cette annotation est d'identifier l'*intention* des utilisateurs par rapport au système. Il s'agit donc d'une annotation *ad hoc* très proche de l'application, sans portée générale.

Trois exemples de dialogue de complexité différente sont présentés dans la table 3.12.

En plus de l'annotation en type d'appels, un deuxième niveau d'annotation est disponible dans ce corpus, il s'agit de l'annotation en entités nommées. Les seules entités annotées sont celles qui ont une utilité pour la résolution de la tâche. Par conséquent leurs définitions sont intimement liées à l'application et au moteur de gestion du dialogue. Deux types d'entités nommées sont utilisées : les entités *génériques* telles que les dates, les montants monétaires ou encore les numéros de téléphone ; les entités *spécifiques* qui n'ont de sens que dans le contexte de l'application HMIHY.

Un exemple d'entité spécifique est l'entité `Which_Bill`[2] qui fait référence à une facture d'un utilisateur. Les supports de cette entité peuvent être :

- « *my January bill* » (« *ma facture de janvier* ») : dans ce cas la facture est identifiée par la période couverte :
- « *my previous statement* » (« *ma facture précédente* ») : facture identifiée de manière relative ;
- « *my bill issued on the second of January* » (« *ma facture datée du 2 janvier* ») : cette fois c'est la date d'émission qui identifie la facture.

Dans tous les cas la valeur associée à l'entité est un numéro d'identifiant de facture.

Cet exemple montre les possibles ambiguïtés d'étiquetage des entités nommées. En effet, le support « *second of January* » peut être soit associé à l'entité `Which_Bill` comme dans l'exemple précédent, soit à l'entité générique date `Item_Date`.

Le tableau 3.13 présente des exemples d'entités nommées avec leur forme support et la valeur extraite.

[1] « *Comment puis-je vous aider ?* »
[2] *quelle facture*

Exemple de dialogue simple
- **System :** How may I help you ?
- **User :** Hello ?
- *Call-type : Hello*
- **System :** Hello, how may I help you ?
- **User :** I have a question.
- *Call-type : Ask(Info)*
- **System :** OK, What is your question ?
- **User :** I would like to know my account balance.
- *Call-type : Request(Account_Balance)*
- **System :** I can help you with that. What is your account number ?
- **User :** ...

Exemple de dialogue avec routage vers un service
- **System :** How may I help you ?
- **User :** I want to pay a bill.
- *System Call-type : Make(Payment)*
- **System :** Do you want to arrange a bill payment ?
- **User :** Yes.
- *System Call-type : Yes*

Exemple de dialogue avec requête non couverte par l'application
- **System :** How may I help you ?
- **User :** Yes somebody just called me from I don't know it's collect or something like that and it had to do with A T and T and when...
- *System Call-type : Explain(Bill)*
- **System :** Do you have a question about a specific charge on your bill ?
- **User :** No
- *System Call-type : No*
- **System :** I'm sorry . How may I help you ?
- **User :** It just said call you guys for help about trying to get through to a long distant number and it's not getting through when I called the long distant ...
- *System Call-type : Other*

TAB. 3.12: *Exemples de dialogues de complexité différente extraits du corpus d'AT&T How May I Help You ?*

étiquette	support	valeur
Item_Amount	« this 22 dollar charge »	22.00
Phone	« 386 5715 area code 201 »	2013865715
Date	« June tenth »	????/06/10
Item_Label	« it says unpaid balance »	unpaid balance
Which_Bill	« most recent statement »	latest

TAB. *3.13:* *Exemples d'entités nommées du corpus HMIHY avec leurs étiquettes, leurs contextes et leurs valeurs.*

3.3.2 Le corpus France Télécom du service vocal 3000

Le service **3000** est le premier service déployé à France Télécom acceptant la parole spontanée non contrainte. Il a été mis en service en Octobre 2005. Ce service permet aux clients de France Télécom d'obtenir des renseignements, d'acheter environ 30 services liés à leur ligne téléphonique, de consulter leur consommation, de payer leur facture et de gérer des services tels que le transfert d'appel ou la messagerie vocale. Le système est fondé sur un module de reconnaissance de la parole continue avec un modèle de langage n-grammes, le module d'interprétation sémantique est le système VERBATEAM fonctionnant en deux étapes : une première étape traduit le message transcrit automatiquement en une série de concepts liés à l'application ; un ensemble de règles manuellement définies (plus de 2600 règles pour le **3000**) compose ces concepts pour produire une interprétation sous la forme prédicat/argument.

Etant données les fonctionnalités de l'application, deux types de dialogue peuvent être distingués : les dialogues d'utilisateurs réguliers (*transit*) et ceux provenant de nouveaux utilisateurs (*autre*).

les dialogues d'utilisateurs réguliers, ayant déjà souscrit aux fonctionnalités du **3000**, telles que la vérification de la consommation ou le transfert d'appel. Ces appels sont automatiquement dirigés vers des sous-dialogues traitant spécifiquement ces tâches. Dans ce cas ce service peut être vu comme un portail d'accès pour utilisateurs abonnés, redirigeant les demandes vers des applications spécialisées, en utilisant éventuellement les profils utilisateurs. Les messages collectés dans ce type de dialogue sont pour la plupart très courts, les utilisateurs réguliers ayant appris à communiquer avec le système se contentent généralement de mots clés pour définir leur requête. Les dialogues sont aussi très courts, entre un et trois tours de dialogue pour atteindre l'application demandée. Ces dialogues sont appelés des dialogues *transit* et représentent 80% des appels au service **3000**.

Les 20% restants sont appelés les dialogues *autre*. Les appels proviennent d'utilisateurs novices ou irréguliers, demandant des informations ou désirant acheter un service. La taille des messages et des dialogues est supérieure à celle des dialogues *transit*, les utilisateurs ne sachant pas toujours comment formaliser leur demande. Par conséquent le taux de disfluences dans ce type de dialogue est aussi assez important. La table 3.14 présente des exemples de requêtes provenant de ces différents types de dia-

Exemple de requêtes *transit*
– désactiver mon transfert
– transfert d'appel
– payer mon facture
– messagerie vocale

Exemple de requêtes *autres* mais couvertes par l'application
– je vous appelle à propos de d' une facture qui n' a pas été réglée et qui a été réglée alors je voudrais avoir quelqu' un pour m' expliquer avec
– oui je voudrais un renseignement qui est très important s'il-vous-plaît
– pourquoi j' ai plus de tonalité
– je voudrais savoir si j' ai reçu la facture que je dois payer ou si je n' ai pas reçu de facture ou si je vous dois de l' que je rentre de l' et je vais y retourner c' est tout ce que je veux savoir mon numéro c' est le je sais plus du tout où j' en suis je suis très fatiguée et je retrouve pas la facture si je dois vous payer ou si je vous ai payée je sais rien du tout alors savoir si vous m' avez envoyé une facture le montant et quand je dois vous la payer

Exemple de requêtes non couvertes par l'application
– je comprends pas pourquoi on me dit j' entends la sonnerie et puis quand je vais pour décrocher il y a il y a plus rien je comprends pas ça à plusieurs fois à plusieurs ça fait ça
– allo je voudrais connaître le numéro d' un d' un pressing qui a changé de propriétaire et et de numéro de téléphone évidemment je ne sais pas si je suis bien reliée

TAB. *3.14:* *Exemples de requêtes de complexité différente extraits du corpus de France Télécom FT3000*

logues.

	autre	**transit**
nb dialogues	350	467
nb messages	1288	717
nb mots	4141	1454
nb moyen de tours de dialogue	3.7	1.5
taille moyenne des messages	3.2	2.0
taux de mots hors-vocabulaire (%)	3.6	1.9
taux de disfluences (%)	2.8	2.1

TAB. *3.15:* *Statistiques décrivant les corpus transit et autre*

Les caractéristiques des deux types de dialogue sont présentées dans la table 3.15. Comme nous pouvons le constater les utilisateurs novices ou irréguliers produisent des messages plus longs, avec plus de disfluences (exprimées ici par les faux départs et les marqueurs de pause) et un taux de mots hors-vocabulaire par rapport au lexique du module de RAP bien supérieur à celui des utilisateurs réguliers.

Une autre caractéristique importante des messages classés comme *autre* est le nombre important de *commentaires* laissés par ces utilisateurs novices. Le terme *commentaire*, dans ce document, représente les énoncés ou portions d'énoncés considérés comme hors du domaine sémantique de l'application. Notamment les utilisateurs novices, lorsqu'ils sont face à une machine, ont tendance à commenter les réactions de la machine à leur requête. Par exemple on trouve des commentaires tels que : « *qu'est ce que je dois dire maintenant* » ; « *j'ai jamais dit ça* » ou bien encore des insultes. Certains messages contiennent uniquement des commentaires, d'autres mélangent commentaires et informations utiles. Comme montré dans la table 3.16, environ 10% des messages et 14% des dialogues *autre* contiennent des commentaires de ce type.

	autre	**transit**
nb dialogues	350	467
nb messages	1288	717
% commentaire par message	10.6	3.3
% commentaire par dialogue	14.3	3.6

TAB. *3.16:* *Distribution des segments commentaires dans les dialogues transit et autre du corpus FT3000*

Ce sont bien évidemment les messages des utilisateurs novices qui seront les plus intéressants à traiter, du point de vue de la recherche en traitement automatique du langage. Notons tout de même qu'il est heureux, du point de vue industriel, que 80% des messages soient des messages *transit* plus simples à traiter de manière robuste !

Chapitre 4

Des mots vers les concepts

Sommaire

Dans la section 2.2.4, j'ai tenté de spécifier quelles étaient les contraintes liées au traitement de messages oraux. A partir de ces contraintes, et de celles provenant des corpus annotés disponibles, nous avons adopté un modèle de représentation du sens fondé sur trois niveaux d'analyse :

1. le premier niveau décrit le sens d'un message par une séquence de *concepts* ; chaque concept est une paire attribut/valeur qui représente une *unité de sens* minimale ayant soit une portée générale (par exemple les entités nommées ou certaines classes de verbes), soit une portée limitée au cadre applicatif dans lequel est construit le modèle (par exemple le concept `which_bill` présenté dans la section 3.3.1 pour le corpus HMIHY) ; cette opération de représentation du sens en séquence de concepts est appelé *décodage conceptuel* ;
2. le deuxième niveau assemble ces unités de sens pour produire une interprétation formelle du message ; cette opération, appelée *composition sémantique*, peut prendre diverses formes en fonction des cadres applicatifs présentés au chapitre 3 : attribution d'une étiquette unique à chaque message dans le corpus HMIHY ; obtention d'une structure prédicat/argument précisant la prochaine action à effec-

tuer dans le corpus FT3000 ou encore production des cadres sémantiques précisant le sens du message avec le formalisme *FrameNet* comme défini dans le projet LUNA et appliqué au corpus MEDIA ;

3. enfin le dernier niveau intègre le contexte de production du message pour enrichir et éventuellement modifier l'interprétation produite ; dans le cadre du dialogue oral ce niveau permet également d'effectuer deux opérations cruciales : la résolution des références entre les tours d'un même dialogue et la segmentation et la caractérisation des segments par rapport à un ensemble d'actes de dialogues.

Nous allons maintenant décrire les méthodes étudiées pour produire automatiquement ces trois niveaux d'analyse, en commençant par l'opération de *décodage conceptuel*.

Pour implémenter un décodeur permettant de produire une séquence de concepts à partir d'un message oral, il convient de spécifier à la fois un formalisme de représentation des concepts avec sa traduction informatique et sa méthode d'analyse, mais aussi le format des données prises en entrée sur lesquelles cette méthode d'analyse sera appliquée : signal de parole, décodage phonétique, graphe de mots, liste de séquences ou encore séquence unique de mots.

La section 2.2.4 a illustré le besoin de robustesse des systèmes de compréhension face aux problèmes liés au traitement de la parole spontanée : disfluences, agrammaticalité, énoncés hors-domaine. Ces phénomènes sont mal modélisés par les systèmes de RAP actuels, chacun d'eux entraînant de nombreuses erreurs de reconnaissance. Pour cette raison il nous paraît important de ne pas laisser au seul module de RAP la tâche de transcription en mots d'un message vocal, mais au contraire d'intégrer les tâches de transcription et de décodage conceptuel au sein d'un même processus, afin de produire en même temps les séquences de mots et les séquences de concepts les représentant.

Cette intégration nécessite cependant une première phase effectuée uniquement par un processus de RAP projetant dans un espace de recherche lexical les paramètres acoustiques extraits du signal de parole. En effet, le traitement de la parole spontanée nécessite l'emploi de modèles de langage robustes de type n-grammes afin d'effectuer un premier décodage limitant l'espace de recherche des transcriptions possibles d'un message à un ensemble d'hypothèses de taille acceptable.

Le processus de décodage proposé, intégrant le décodage en mots et concepts, fonctionne en quatre étapes :

1. tout d'abord l'espace des transcriptions en mots possibles d'un message est réduit, dans la première phase du processus de RAP, à l'ensemble des séquences de mots candidates, chaque mot étant valué avec son score acoustique et son score linguistique ;
2. chaque séquence de mots est projetée vers un ensemble de séquences de concepts qui représentent l'ensemble des interprétations possibles, sous forme conceptuelle, de la séquence considérée ;
3. chaque paire (séquence de mots W / séquence de concepts Γ) est évaluée grâce à la probabilité jointe $P(W, \Gamma)$; une liste de n-meilleures hypothèses triées selon cette probabilité est produite, cette liste est *structurée* par rapport aux différentes interprétations conceptuelles du message considéré ;

4. enfin la dernière étape est constituée par le module de décision qui va choisir parmi les n-meilleures hypothèses (W, Γ) en estimant la probabilité qu'une paire (W, Γ) soit correcte étant donné un ensemble de mesures de confiance. Ces mesures de confiance sont obtenues sur les différents niveaux du décodage mais aussi sur le contexte de production du message.

4.1 Choix de l'espace de recherche

L'ensemble des séquences de mots candidates produit par la première phase de décodage peut être représenté sous plusieurs formes dont les trois principales sont : listes de n-meilleures hypothèses, graphes de mots et réseaux de confusion. Les deux premières sont directement issues du décodeur du module de RAP : les n-meilleures solutions sont produites en énumérant les n-meilleurs chemins de l'espace de décodage ; les graphes de mots sont une réduction autour du meilleur chemin de cet espace de recherche. L'algorithme de production d'un graphe de mots dépend fortement de l'algorithme de décodage utilisé par le module de RAP (voir par exemple dans (Ljolje et al., 1999) une méthode efficace de production de graphes de mots).

Les listes de n-meilleures hypothèses de séquences de mots ont l'avantage d'être directement utilisables par n'importe quel module de traitement de la langue acceptant du texte en entrée : chaque hypothèse est considérée comme un texte différent. Leur principal défaut est que la différence entre chaque séquence se fait au niveau des mots (un mot au minimum), sans aucune distinction de l'importance de ceux-ci pour l'extraction du sens. De plus les mots inconnus ou les disfluences telles que les hésitations et les reprises se traduisent toujours pour le décodeur par une zone d'instabilité où de nombreux mots peuvent être soit insérés, soit supprimés.

Par exemple le message : « alors euh le le deux avril » pourra être transcrit par les séquences suivantes : *alors le le le deux avril*, *alors euh le deux avril*, *alors euh le le deux avril*, *alors le euh avril*, ...
Si un message contient plusieurs zones d'instabilité, le nombre de chemins potentiels étant exponentiel, il faudra garder un nombre n de séquences potentiellement énorme afin de garantir la présence d'hypothèses différentes au niveau du *sens*, plutôt que par rapport à des mots outils souvent peu informatifs.

Les graphes de mots sont des graphes orientés, connexes et acycliques. Le nombre de séquences de mots possibles qu'ils contiennent dépend de l'algorithme utilisé pour les générer et du facteur d'élagage choisi. Ce facteur d'élagage permet de contrôler la taille de l'espace de recherche. Quel que soit le facteur choisi, le nombre de chemins possibles est toujours très important, les zones d'instabilité décrites précédemment produisent des portions de graphe très denses. Cette richesse, en terme de nombre de chemins, garantit d'obtenir des séquences de mots produisant des interprétations différentes du message traité. Ces graphes peuvent être représentés par des automates à états finis (appelés FSM pour *Finite State Machine* dans ce document) et manipulés grâce aux très nombreuses opérations définies sur les automates, comme nous le ver-

rons dans le paragraphe suivant. Les graphes ont cependant deux inconvénients : d'une part ils nécessitent de modifier les modules destinés à traiter du texte en entrée ; d'autre part leur taille est souvent un obstacle à leur traitement dans les applications ou les contraintes sur le temps de traitement sont fortes, comme par exemple les applications de dialogue.

Les réseaux de confusion sont obtenus à partir des graphes de mots en appliquant un certain nombre d'heuristiques pour obtenir, pour chaque graphe, une chaîne d'ensembles de mots concurrents. Le score de chaque mot est la somme des probabilités *a posteriori* des transitions portant sur ce mot dans la zone de concurrence correspondante. Une transition vide symbolisant l'omission est éventuellement ajoutée pour assurer le complément à 1 des scores dans chaque zone. Les réseaux de confusion ont été proposés par (Mangu et al., 2000), un algorithme alternatif et son application pour deux tâches de compréhension sont présentés dans (Hakkani-Tür et al., 2006).

Les réseaux de confusion ont plusieurs avantages : d'une part leur taille est très réduite par rapport au graphe desquels ils sont extraits, sans nécessairement perdre beaucoup de chemins ; d'autre part ils sont faciles à découper, et le score de chaque mot est directement interprétable comme une mesure de confiance. Par contre ils ont l'inconvénient de dépendre fortement des heuristiques choisies pour grouper les mots au sein d'un même ensemble et pour déterminer le nombre d'ensembles. Le nombre de chemins possibles d'un réseau de confusion est aussi potentiellement énorme, supérieur à ce que l'on peut trouver dans le graphe lui correspondant. En effet chaque mot d'un ensemble peut être suivi par n'importe quel mot de l'ensemble suivant, y compris la transition vide. Il existe ainsi de très nombreux chemins qui sont ajoutés lors de la production du réseau alors qu'ils n'étaient pas dans l'espace de recherche de départ.

Dans le modèle présenté dans ce chapitre nous utilisons les graphes mots comme espace de recherche à la frontière entre le module de RAP et le processus de compréhension, ces graphes de mots sont représentés sous la forme d'automates à états finis (FSM). Cependant nous verrons dans le chapitre 6 une stratégie de décodage qui utilise à la fois des listes de n-meilleures hypothèses, des réseaux de confusion et des graphes.

4.2 Représentation des concepts par des automates à états finis

Nous utilisons des automates à états finis (FSM) pour représenter les concepts dans le modèle présenté dans ce chapitre. Ce choix est motivé par les arguments suivants :

- dans la mesure où les concepts sont exprimés avec des séquences de mots très courtes et peu complexes, il n'est pas nécessaire d'utiliser un formalisme de représentation plus riche que les grammaires régulières pour les représenter ;
- en utilisant les FSM à la fois pour les graphes de mots issus de la RAP et les concepts à retrouver, nous profitons de ce formalisme unique et de toutes les opérations qui y sont associées pour rechercher ces concepts directement dans les graphes ;
- la séquence de mots W n'étant pas fixée, nous désirons chercher simultanément

W et Γ, ce qui implique qu'il est plus facile d'utiliser pour cela un modèle génératif estimant $P(W,\Gamma)$ plutôt qu'un modèle discriminant cherchant la probabilité *a posteriori* sur l'hypothèse W : $P(\Gamma|W)$;
- étant donné que la plupart des concepts contiennent une valeur en plus de l'attribut les définissant, il est de toute manière indispensable de réaliser une analyse, pour chaque concept, de sa chaîne de mots support afin d'en extraire la valeur ; l'extraction de valeur ne pouvant être vue comme un processus de classification uniquement dans le cas d'attributs prenant leurs valeurs dans un ensemble fermé (et de petite taille) ;
- enfin on veut pouvoir rajouter facilement des connaissances explicites ou des modèles appris sur d'autre données au modèle de représentation des concepts d'une application particulière ; les FSM se prêtent facilement à ce genre d'opérations comme nous allons le voir dans la prochaine section.

4.2.1 Obtenir des grammaires de concepts à partir de corpus

Certains concepts génériques, tels que les dates ou encore les numéros de téléphone, sont communs à de nombreuses applications, et obéissent à des règles de construction très facilement modélisables par des grammaires régulières. A l'inverse pour des concepts plus spécifiques d'une application particulière il est intéressant, sous réserve de la disponibilité d'un corpus transcrit, de construire automatiquement ces grammaires régulières à partir des exemples du corpus. Les annotations nécessaires sont les transcriptions en mots et les balises indiquant le début et la fin de la forme support de chaque concept. Lorsque l'annotation en segments des concepts n'est pas disponible, une méthode automatique utilisant uniquement l'indication de la présence ou l'absence d'un concept dans un message peut être utilisé pour obtenir ces formes supports, comme présenté dans (Béchet et al., 2002).

Plusieurs approches ont été proposées pour obtenir automatiquement des grammaires à partir de corpus d'exemples (Ron et al., 1998; Carrasco et Oncina, 1999; Stolcke et Omohundro, 1994). Ajouter un nouvel exemple à une grammaire dont l'axiome est le symbole non-terminal S consiste à ajouter une nouvelle règle de production pour S qui couvre précisément l'exemple. Puis un symbole non terminal est ajouté pour chaque symbole terminal sur la partie droite de cette nouvelle règle.

Les algorithmes diffèrent sur la stratégie choisie pour regrouper les différents symboles non-terminaux : si aucune fusion n'est faite, la grammaire se contente de modéliser précisément le corpus d'apprentissage, sans aucune généralisation. Si trop de symboles non-terminaux sont fusionnés, le risque d'accepter des phrases incohérentes augmente. Dans les applications de ce modèle présentées dans ce document, sur les corpus PLANRESTO, HMIHY ou MEDIA, les seuls regroupements effectués sont ceux correspondant à des classes de non-terminaux bien identifiés : les nombres, les dates et certaines classes de noms propres telles que les noms de lieux. Ce regroupement, très simpliste, est justifié par la raison d'être de ces grammaires : permettre l'analyse de la forme support d'un concept en vue de l'extraction de sa valeur. Les grammaires issues

du corpus doivent ainsi être totalement compatibles avec les règles de production des valeurs présentées dans la section 4.4.

Par exemple, pour l'application HMIHY, nous avons utilisé (Bechet et al., 2004) les généralisations suivantes :

- chaque chiffre (0 to 9) est remplacé par le symbole `$digitA;`
- les entiers naturels de 10 à 19 sont remplacés par le symbole `$digitB;`
- chaque multiple de 10, différent de 10, (20, 30, 40, ...90) est remplacé par le symbole `$digitC;`
- les ordinaux sont remplacés par `$ord;`
- chaque jour de la semaine est remplacé par `$day;`
- chaque mois est remplacé pat `$month;`

Par exemple, les six supports du concept `Item_Amount` de l'application de HMIHY

```
charged for two ninety five ; charging me a dollar sixty five
charged a dollar sixty ; charges of thirty dollars
charge of eleven dollars and sixty three cents
charged to me for four dollars and forty eight cents
```

sont remplacés par :

```
charged for $digitA $digitC $digitA ;
charging me a dollar $digitC $digitA
charged a dollar $digitC ; charges of $digitC dollars
charge of $digitB dollars and $digitC $digitA cents
charged to me for $digitA dollars and $digitC $digitA cents
```

Toutes les règles de grammaire obtenues pour un concept donné sont représentées par un FSM, comme illustré sur la figure 4.1.

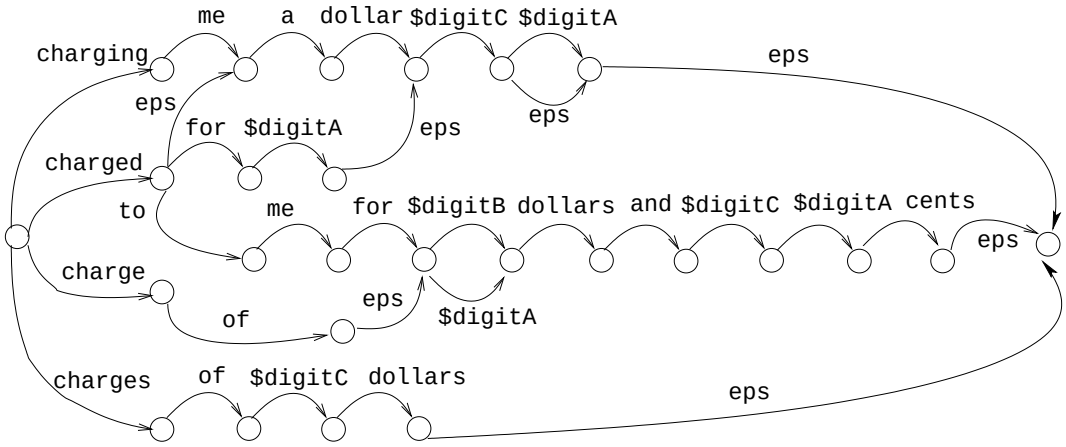


FIG. 4.1: *FSM obtenu à partir du corpus HMIHY pour le concept Item_Amount*

Les grammaires obtenues ne sont pas stochastiques. Dans notre stratégie c'est le rôle du modèle de décodage mots/concepts d'estimer la probabilité jointe d'une séquence de mots et d'une séquence de concepts, comme cela sera présenté dans la section 4.5. Ce modèle sera appris sur un corpus spécifique à l'application visée. Pour les grammaires

de concepts, le fait qu'elles ne soient pas valuées permet de les enrichir très facilement avec des règles définies manuellement ou obtenues à partir d'autres corpus provenant d'autres applications. C'est particulièrement le cas pour les concepts génériques comme les dates, ou pour certains concepts obéissant à une grammaire particulière telle que celle des numéros de téléphone, facile à décrire manuellement de manière exhaustive.

4.3 Projection d'un graphe de mots vers un graphe de concepts

4.3.1 Représentation et manipulation des graphes

Toutes les opérations faites avec des FSM dans cette étude utilisent la bibliothèque d'AT&T FSM Library (Mohri et al., 1997). En suivant les notations utilisées dans (Mohri et al., 2002), les FSM accepteurs et transducteurs utilisés sont définis par la structure algébrique de *semi-anneau*. Un semi-anneau K est constitué d'un ensemble $\mathbb{K}$ avec une opération associative et commutative $\oplus$, une opération associative $\otimes$, avec deux éléments d'identité : $\overline{0}$ pour $\oplus$ et $\overline{1}$ pour $\otimes$.
On a : $K = (\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$.

Les graphes de mots issus du module de RAP où les poids associés aux transitions représentent les probabilités P des mots exprimés en *-logprob* ($-\log(P)$) correspondent au semi-anneau *-log* : $(\mathbb{R}, +, ., 0, 1)$. Pour trouver le plus court chemin dans un tel graphe on utilise le semi-anneau $\mathbb{R}_{min}$ appelé aussi *semi-anneau tropical* correspondant à : $(\mathbb{R}_+ \cup \{\infty\}, min, +, \infty, 0)$.

Les accepteurs et les transducteurs sont définis de la manière suivante :

Soit Σ un alphabet de symboles d'entrée ; Δ un alphabet de symboles de sortie ; ϵ un symbole vide ; Q un ensemble d'états (avec I=état initial et F=états finaux) ; $\mathbb{K}$ un semi-anneau ; E un ensemble de transitions définies tel que : $E \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times (\Delta \cup \{\epsilon\}) \times \mathbb{K} \times Q$; w un fonction de coût : $w : Q \to \mathbb{K}$.

Si $Path(R_1, x, R_2)$ est un ensemble de chemins de $R_1 \subseteq Q$ vers $R_2 \subseteq Q$ avec le label d'entrée x et $Path(R_1, x, y, R_2)$ un ensemble de chemins de $Path(R_1, x, R_2)$ avec le label de sortie y, alors :

– Accepteur $A = (\Sigma, Q, I, F, E)$ avec pour tout $x \in \Sigma$:

$$[A](x) = \bigoplus_{\pi \in Path(I,x,F)} w[\pi] \tag{4.1}$$

– Transducteur $T = (\Sigma, \Delta, Q, I, F, E)$ avec pour tout $x \in \Sigma^*, y \in \Delta^*$:

$$[T](x,y) = \bigoplus_{\pi \in Path(I,x,y,F)} w[\pi] \tag{4.2}$$

et $w[\pi] = w[t_1] \otimes w[t_2] \otimes \ldots \otimes w[t_n]$ pour un chemin π composé des transitions suivantes $t_1, t_2, \ldots, t_n$.

Nous utilisons les opérations suivantes sur de tels accepteurs et transducteurs :
- Composition : $[T_1 \circ T_2](x,y) = \bigoplus_z [T_1](x,z) \otimes [T_2](z,y)$
- Intersection : $[A_1 \cap A_2](x) = [A_1](x) \otimes [A_2](x)$
- Différence : $[A_1 - A_2](x) = [A_1 \cap \overline{A_2}](x)$
- Projection : $[A](x) = \bigoplus_y [T](x,y)$ *and* $[A](y) = \bigoplus_x [T](x,y)$

4.3.2 Principe de décodage

Chaque concept $\gamma_k \in \Gamma$ est représenté par un accepteur FSM (A_k pour le concept γ_k). Les chaînes de mots n'appartenant à aucun concept sont reconnues par un modèle *mange-mot* ou *filler* appelé A_F.

Tous ces accepteurs sont transformés en transducteurs prenant les mots en entrée et produisant les symboles de début (`start`) et fin (`end`) de concepts. Les accepteurs A_k deviennent les transducteurs T_k avec la première transition émettant le symbole $<\gamma_k>$ et la dernière transition le symbole $</\gamma_k>$. De manière similaire le modèle *filler* est représenté par le transducteur T_{BK} qui émet les symboles $< BAK >$ et $< /BAK >$. Finalement tous ces transducteurs sont groupés ensemble dans un seul modèle appelé $T_{concept}$ comme présenté dans la figure 4.2.

Le principe de décodage et la production d'un graphe de concepts à partir d'un graphe de mots est présenté ici sur un exemple simple. Le graphe de mots produit par le module de RAP à partir d'un message vocal est représenté par un accepteur G_W. Le semi-anneau utilisé est le semi-anneau *log* et la fonction de coût $w(\pi)$ correspond au $-\log$ de la probabilité $P(W|Y)$, probabilité d'avoir la séquence de mots W (représentée par le chemin π) à partir de la séquence d'observations acoustiques Y. Un exemple de FSM G_W est donné dans la figure 4.3.

G_W est *composé* avec le transducteur $T_{Concept}$ de manière à obtenir le transducteur mot/concept T_{WC} : $T_{WC} = G_W \circ T_{Concept}$, illustré par la figure 4.4.

Dans cet exemple, extrait du corpus PLANRESTO, on considère quatre types de concepts : les modifieurs `NEAR` (pour un lieu), `LESS` (pour un montant) et les entités nommées `LOC` pour les lieux et `AMOUNT` pour les quantités monétaires.

```
NEAR   = pas loin du metro  (near metro)
NEAR   = pas loin du $NAME (near $NAME)
LOC    = $NAME
LOC    = metro $NAME
AMOUNT = $NUMBER euros
LESS   = moins de $NUMBER (less than $NUMBER)
```

avec `$NAME` pouvant être n'importe quel nom propre et `$NUMBER` n'importe quelle expression numérique.

Un chemin $Path(I,x,y,F)$ dans T_{WC} (avec I l'état initial et F l'état final de T_{WC}) est soit une chaîne de mots si on considère les symboles d'entrée x, soit une chaîne de

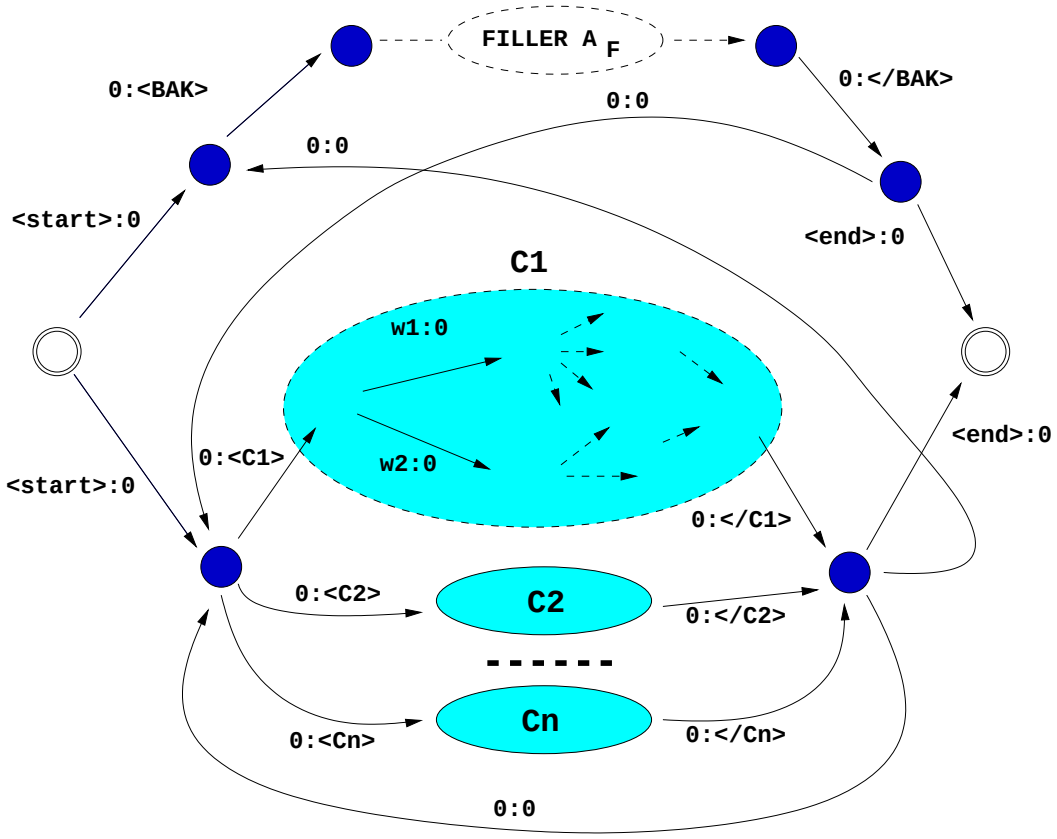


FIG. 4.2: *Exemple de transducteur de mots vers les concepts* $T_{concept}$

concepts si on considère les symboles de sortie y. De manière à obtenir le graphe de concepts le plus compact contenant toutes les séquences de concepts présentes dans G_W, on projette T_{WC} sur les symboles de sortie, puis on déterminise et minimise le FSM obtenu. L'accepteur obtenu est appelé G_C. Etant donné que cette opération est effectuée dans le semi-anneau log, le coût d'un chemin $Path(I, y, F)$ est la somme de tous les coûts des chemins produisant y dans T_{WC} :

$$[G_C](y) = \bigoplus_x [T_{WC}](x, y) \tag{4.3}$$

Un exemple de FSM G_C, obtenu à partir du transducteur T_{WC} de la figure 4.4 est donné sur la figure 4.5. Par souci de clarté, les tags de fin des concepts sont omis. Les états doublement cerclés sont des états terminaux.

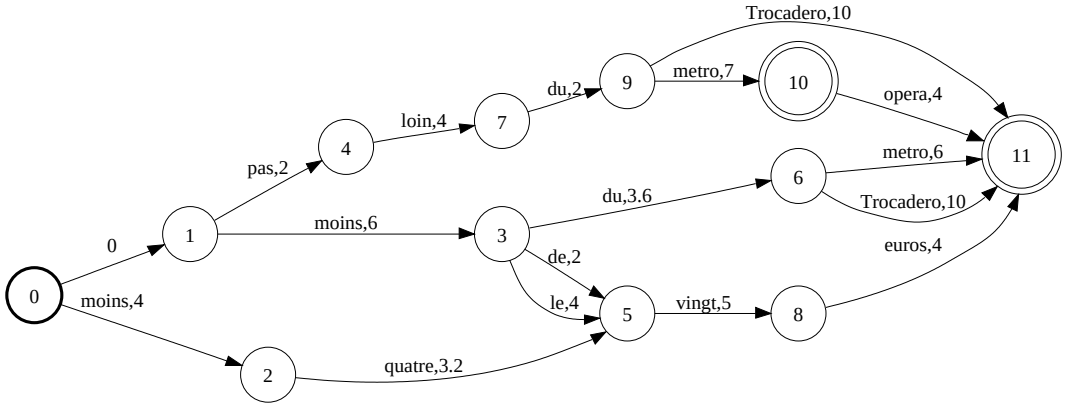


FIG. *4.3:* *Exemple de graphe de mots G_W produit par un module de RAP*

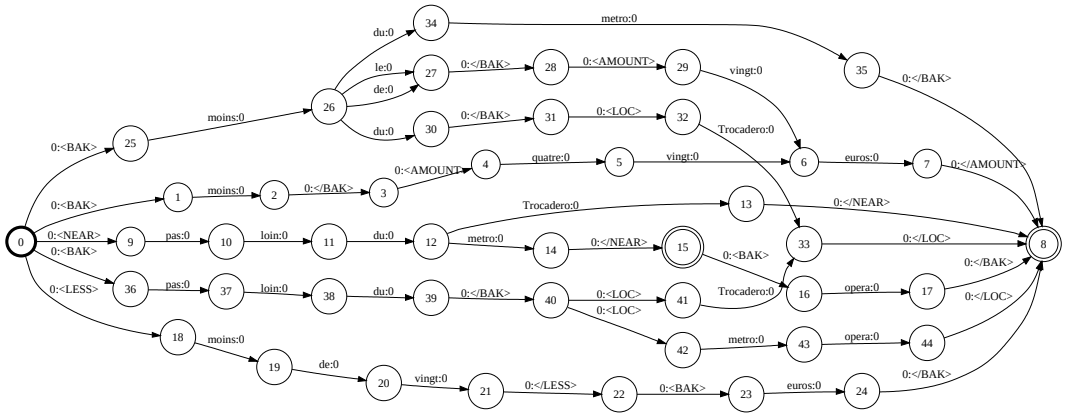


FIG. *4.4:* *Exemple de transducteur T_{WC} correspondant à la composition d'un accepteur représentant un graphe de mots et le transducteur mot/concept $T_{Concept}$*

4.4 Extraction d'une liste de *n*-meilleures valeurs pour chaque séquence de concepts

La valeur d'un concept doit être extraite de sa forme support et normalisée. C'est sur cette forme normalisée que l'évaluation du module de compréhension pourra être faite, plus que sur les mots de la forme support. En effet, c'est cette forme normalisée qui aura un impact sur les performances globales d'un système utilisant le module de compréhension développé. Par exemple, la valeur associé à la forme support « *bill issued on November 12th 2001* » du concept `which_bill` est *2001/11/12*, et du moment que cette valeur et cet attribut sont reconnus, peu importe que la transcription en mots ait produit « *bill of the November 12th 2001* » ou « *issue the November 12th of 2001* ».

Extraire une valeur à partir d'une chaîne de mots n'est pas toujours une opération facile à cause des éventuelles ambiguïtés de projection. Par exemple, dans le corpus HMIHY, un numéro de téléphone tel que *220 386 1200* peut être lu de la manière suivante : *two twenty three eight six twelve hundred*. Cette chaîne peut à son tour être traduite

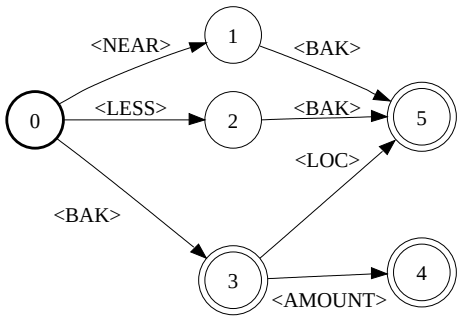


FIG. 4.5: *Exemple d'accepteur G_C obtenu en projetant le transducteur T_{WC} sur les symboles de sortie*

par les séquences de chiffres suivantes :
`2203861200  223861200 22038612100  2238612100` .....

Cette étape de projection peut être faite avec des règles *ad hoc*, spécifiques à chaque application. On peut aussi utiliser une méthode à base d'apprentissage automatique si suffisamment de corpus est disponible (Raymond et al., 2006; Bonneau-Maynard et Lefevre, 2005). Dans ces travaux nous utilisons principalement un mécanisme de transduction permettant de lever les ambiguïtés de projection. Par exemple, cette transduction appliquée au message précédent produit :
`two->2 twenty->20 three->3 eight->8 six->6 twelve->12 hundred->00`

Pour les grammaires écrites manuellement, cette transduction est spécifiée en ajoutant à chaque symbole terminal le format du symbole de sortie qui devra être émis. Par exemple, la transduction précédente est spécifiée par les règles :
```
<PHONE> -> $digitA/$digit1 $digitC/$digit2 $digitA/$digit1
           $digitA/$digit1 $digitA/$digit1 $digitB/$digit2 hundred/00
```

avec `$digit1` correspondant au premier chiffre du symbole d'entrée, `$digit2` aux deux premiers chiffres du même symbole, et `00` à la séquence de chiffre *00*.

Ces règles de transduction sont appliquées au transducteur mot/concept T_{WC} présenté précédemment, toujours par une opération de composition. L'extraction des valeurs devient ainsi une opération simultanée à la phase de détection des attributs des concepts : une fois qu'une séquence de mots est acceptée par une grammaire de concept, la mise en rapport des symboles d'entrée et de sortie sur le transducteur résultat permet d'enlever toutes les ambiguïtés de projection.

Ce procédé est illustré par la figure 4.6. *FSM1* est le transducteur correspondant au graphe de mots produit par le module de RAP ; *FSM2* est une grammaire représentant la transduction entre une forme support telle que «*sixty four ten charge*» et sa valeur *64.10*. A partir de *FSM1*, les valeurs suivantes peuvent être extraites :
`64.10 64.00 60.10 60.00 60.04 4.10 4.00 10.00 0.64 0.60 0.04 0.10`
Mais après l'opération de composition entre les deux FSM, la seule transduction possible est (avec *eps* représentant la transition vide *epsilon*) :
`sixty->$digit1 four->$digit1 eps->. ten->$digit2 charge->eps`

Ce qui signifie que pour produire une valeur nous devons prendre le premier chiffre de *sixty*, le premier chiffre de *four*, ajouter le symbole ., prendre les 2 premiers chiffres de *ten* et finalement ignorer le mot *charge*. Nous obtenons uniquement une projection possible, 64.10, à la place des douze valeurs possibles avant transduction.

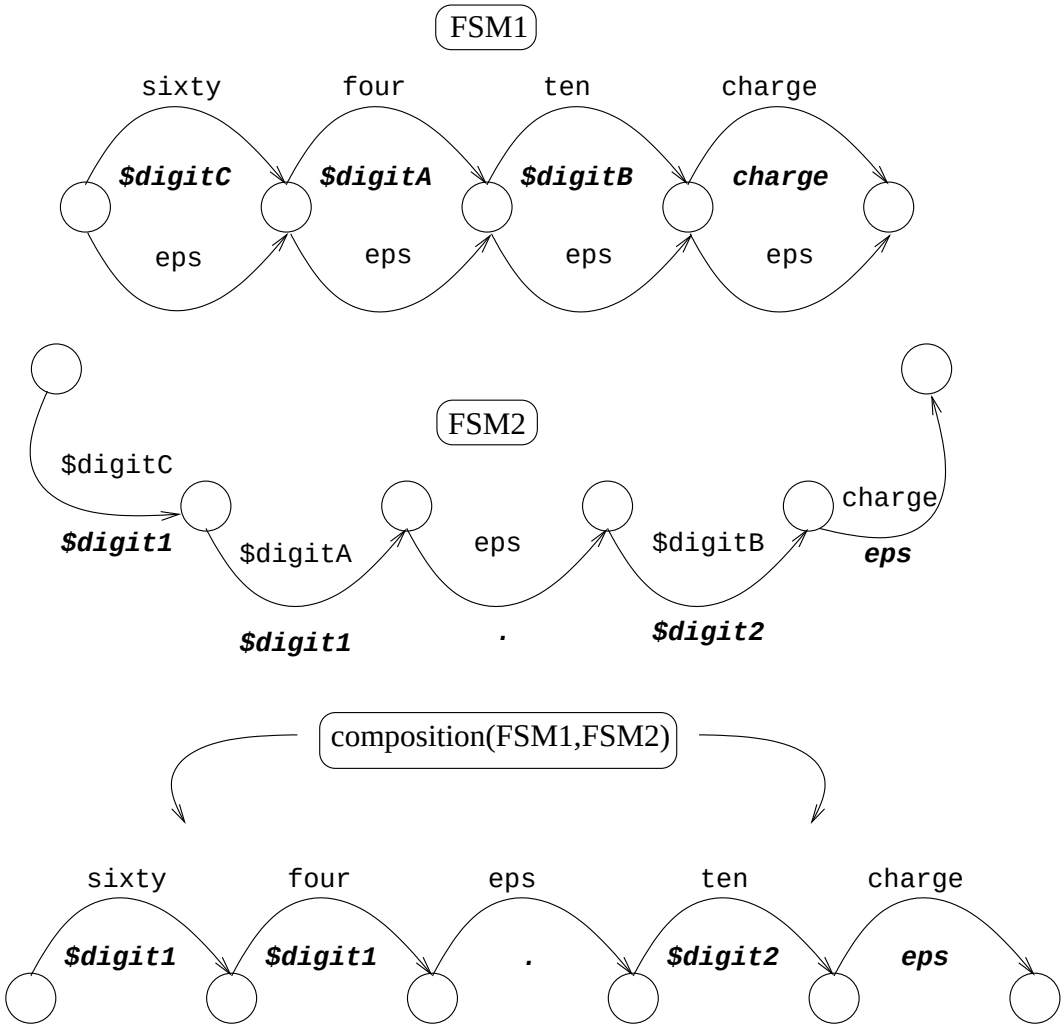


FIG. 4.6: *Exemple de production de valeurs à partir d'une chaîne de mots supports ambigus*

Un des avantages principaux de cette approche est la possibilité de générer directement une liste de *n*-meilleures solutions sur les *valeurs* associées aux concepts. En effet, chaque chemin différent, une fois les transitions vides supprimées dans le transducteur résultat, correspond à une nouvelle valeur.

Enumérer les *n*-meilleurs chemins produisant un concept correspond à énumérer ses *n*-meilleures valeurs. Ceci est un avantage par rapport à énumérer les différents chemins dans un simple graphe de mots, la plupart d'entre eux ne différent que de quelques mots outils sans incidence sur le sens.

Cette extraction des *n*-meilleures valeurs est particulièrement intéressante dans un

cadre de dialogue dans la mesure où l'historique et le contexte du dialogue peuvent permettre de filtrer les listes obtenues par rapport à des contraintes sémantiques contextuelles (par exemple données clients ou espaces de valeurs possibles). Par exemple, (Rahim et al., 2001) montre que filtrer des listes d'hypothèses contenant des numéros de téléphones en utilisant leur pertinence par rapport à un annuaire attestant des numéros valides permet d'effectuer un filtrage très efficace des solutions potentielles : la reconnaissance des numéros de téléphone appartenant à cet annuaire atteint les 94.5% (tout en représentant 61% des hypothèses) à comparer au taux de reconnaissance de 45% pour les numéros hors annuaire.

4.5 Un modèle de langage pour les mots et les concepts

La méthode proposée dans cette étude intègre le processus de décodage conceptuel dans le processus de transcription de parole en remplaçant le modèle de langage par un modèle prédisant à la fois les mots et les concepts. Ce modèle est présenté dans les paragraphes suivants.

4.5.1 Choix de la meilleure séquence de concepts

Le modèle théorique utilisé est le suivant :

soit A la séquence d'observation acoustique représentant le message oral M ;

soit $C = c_1 \dots c_n$ une séquence de concepts c_i issus de M ;

soit $W = w_1 \dots w_k$ une suite de mots w_i supports d'une séquence C ;

Le modèle de décodage conceptuel proposé est un modèle de Markov caché (*Hidden Markov Model* ou *HMM*) à deux niveaux où, sur le premier niveau, les états cachés sont les concepts q_i et les symboles générés sont les mots w_i ; sur le deuxième niveau les états cachés sont les w_i et les symboles générés sont la séquence d'observations acoustiques A décrivant M.

Le choix de la meilleure séquence $\hat{C}$ du message M est fait avec la règle de décision du *Maximum A Posteriori* selon la formule :

$$\begin{aligned} \hat{C} = & \underset{C}{argmax}\ P(C|A) = \underset{C}{argmax} \sum_{W} P(W,C|A) \\ = & \underset{C}{argmax} \sum_{W} P(A|W,C)P(W,C) \approx \underset{C,W}{argmax}\ P(A|W)P(W)P(C|W) \end{aligned} \tag{4.4}$$

On remplace la somme par un *Max* et la probabilité $P(A|W,C)$ par $P(A|W)$ en considérant que la séquence d'observation A ne dépend que des mots W prononcés et pas de la séquence de concepts C.

Avec ce modèle la meilleure séquence $\hat{C}$ est obtenue avec trois sources de probabilités :

- la probabilité $P(A|W)$ donnée par les modèles acoustiques du module de RAP à la suite de mots $W = w_1 \dots w_k$;
- la probabilité $P(W)$ donnée par le modèle de langage du module de RAP ;
- la probabilité $P(C|W)$ du modèle de décodage conceptuel représentant la probabilité d'une séquence de concepts C étant donnée la séquence de mots W.

De manière similaire à ce qui est fait dans les systèmes de RAP pour la tâche de transcription, un facteur d'ajustement est nécessaire entre ces différentes sources de probabilités à cause notamment de la dynamique très différente des probabilités acoustiques et linguistiques. Ces facteurs, estimés sur un corpus de développement en cherchant à minimiser le taux d'erreur sur les concepts, permettent de faire porter un poids plus ou moins grand sur l'un ou l'autre des modèles. La formule 4.4 devient :

$$\hat{C} = \underset{C,W}{argmax} P(A|W)P(W)^{\alpha}P(C|W)^{\beta} \tag{4.5}$$

Comme la probabilité $P(C|W)$ n'est pas directement estimable, on l'inverse pour obtenir :

$$\begin{aligned} \hat{C} &= \underset{C,W}{argmax}\ P(A|W)P(W)^{\alpha}P(W,C)^{\beta}P(W)^{-\beta} \\ &= \underset{C,W}{argmax}\ P(A|W)P(W)^{\alpha-\beta}P(W,C)^{\beta} \end{aligned} \tag{4.6}$$

Le facteur $\alpha - \beta$ est le facteur d'ajustement du modèle de langage seul, le facteur β est le facteur d'ajustement du modèle conceptuel.

Les probabilités $P(A|W)$ et $P(W)$ sont données par les modèles de RAP. L'estimation de la probabilité $P(W,C)$ est présentée dans la prochaine section.

4.5.2 Modèle de langage conceptuel $P(W,C)$

Chaque état du HMM représentant un concept q_i a la possibilité d'émettre un segment composé de plusieurs observations w_j. Il faut donc modéliser la longueur d_i (en nombre d'observations) de chaque concept q_i : $D = d_1 \dots d_n$ pour $C = q_1 \dots q_n$.

Pour la séquence de concepts $C = q_1 \dots q_n$ et la chaîne de mots $W = w_1 \dots w_k$ nous avons :

$$P(W,C) = \sum_{D} P(W,D,C) \tag{4.7}$$

Ce problème s'apparente à un problème de segmentation de texte. Pour cela nous allons adopter la démarche proposée par *Ramshaw* et *Marcus* (Ramshaw et Marcus, 1995)

pour l'analyse de surface (ou *chunking*) qui projette la problématique de la segmentation vers une problématique d'étiquetage. Une étiquette est associée à chaque mot, cette étiquette précise le type de segment dans lequel se trouve le mot ainsi qu'une information binaire sur la position du mot dans le segment : B pour le premier mot du segment (*Begin*) et I pour un mot à l'intérieur du segment (*Inside*). Les mots ne faisant partie d'aucun segment reçoivent l'étiquette O (*Outside*). Ce modèle est souvent appelé le modèle *IOB*.

Pour cela nous notons $(D, C) = (d_1, q_1)\ (d_2, q_2) \ldots (d_n, q_n)$, et chaque couple (d_j, q_j) est remplacé par une séquence de d_j symboles t tel que :

$$(d_j, q_j) = t_1\ t_2 \ldots t_{d_j}$$

Pour chaque (d, q), les symboles t_x ont comme valeur :

$$t_x = \begin{cases} q^b & \text{si } x = 1 \text{ et que } q \neq null \\ q^i & \text{si } x > 1 \text{ et que } q \neq null \\ null & \text{si } q = null \end{cases}$$

Le symbole q^b représentant l'emplacement du premier mot exprimant le concept q (avec b pour *begin*), l'emplacement des autres mots est exprimé par le symbole q^i (avec i pour *inside*). La formule 4.7 devient :

$$\begin{aligned} P(W, C) &= \sum_D P(w_1 \ldots w_k, (d_1, q_1) \ldots (d_n, q_n)) \\ &= \sum_D P(w_{1,k}, t_{1,k}) \approx \underset{D}{argmax} P(w_{1,k}, t_{1,k}) \end{aligned} \tag{4.8}$$

avec $w_{1,k}$ représentant la séquence de k mots $w_1 \ldots w_k$ et $t_{1,k}$ la séquence de k étiquettes $t_1 \ldots t_k$.

Par exemple, pour la séquence de concepts :

$$C = \{q_1 = \texttt{reponse}\ ,\ q_2 = \texttt{localisationVille}\ ,\ q_3 = \texttt{null}\}$$

et la séquence d'observations :

$$W = \{w_1 = \texttt{oui}\ ,\ w_2 = \texttt{à}\ ,\ w_3 = \texttt{Marseille}\ ,\ w_4 = \texttt{bon}\}$$

nous avons (avec les segmentations indiquées après chaque (D, C)) :

$$(D, C) = \begin{cases} (1, q_1), (1, q_2), (2, q_3) = q_1^b\ q_2^b\ q_3^b\ q_3^i \ \rightarrow\ \texttt{[oui][à][Marseille bon]} \\ (1, q_1), (2, q_2), (1, q_3) = q_1^b\ q_2^b\ q_2^i\ q_3^b \ \rightarrow\ \texttt{[oui][à Marseille][bon]} \\ (2, q_1), (1, q_2), (1, q_3) = q_1^b\ q_1^i\ q_2^b\ q_3^b \ \rightarrow\ \texttt{[oui à][Marseille][bon]} \end{cases}$$

d'où :

$$P(W,C) = max \begin{cases} P(w_1\ w_2\ w_3\ w_4\ ,\ q_1^b\ q_2^b\ q_3^b\ q_3^i) \\ P(w_1\ w_2\ w_3\ w_4\ ,\ q_1^b\ q_2^b\ q_2^i\ q_3^b) \\ P(w_1\ w_2\ w_3\ w_4\ ,\ q_1^b\ q_1^i\ q_2^b\ q_3^b) \end{cases}$$

L'étiquetage correct étant :

$w_{1,4} =$ `oui à Marseille bon`

$t_{1,4} =$ `reponse`^{b} `localisationVille`^{b} `localisationVille`^{i} `null`

L'estimation de $P(W,C)$ se ramène donc à une tâche d'étiquetage où chaque observation w_i reçoit un label t_i correspondant au concept qu'il représente et à sa position à l'intérieur de celui-ci. Ce processus est identique à la problématique des étiqueteurs probabilistes, telle qu'on peut la trouver dans (Charniak et al., 1993). En définissant de manière adéquate des termes tels que $t_{1,0}$, ainsi que leurs probabilités, on obtient :

$$P(w_{1,k}, t_{1,k}) = \prod_{i=1}^{n} P(t_i|t_{1,i-1}, w_{1,i-1}) P(w_i|t_{1,i}, w_{1,i-1}) \tag{4.9}$$

De manière à pouvoir estimer ces probabilités, nous faisons les hypothèses de Markov suivantes :

$$\begin{aligned} P(t_i|t_{1,i-1}, w_{1,i-1}) &= P(t_i|t_{i-2,i-1}, w_{i-2,i-1}) \\ P(w_i|t_{1,i}, w_{1,i-1}) &= P(w_i|t_{i-2,i}, w_{i-2,i-1}) \end{aligned}$$

Ainsi nous faisons l'hypothèse que l'étiquette t_i ne dépend que des deux mots et étiquettes précédents. De même le mot w_i ne dépend que des deux mots et étiquettes précédents ainsi que de la connaissance de son étiquette t_i. Nous obtenons l'équation suivante :

$$P(t_{1,n}, w_{1,n}) = \prod_{i=1}^{n} P(t_i|t_{i-2,i-1}, w_{i-2,i-1}) P(w_i|t_{i-2,i}, w_{i-2,i-1}) \tag{4.10}$$

Ce modèle de langage peut être appris directement sur un corpus d'apprentissage étiqueté avec les symboles t_i avec le critère du maximum de vraisemblance.

Chapitre 5

Des concepts vers l'interprétation

Sommaire

Comme décrit dans le chapitre précédent, une fois que les unités de sens ou *concepts* sont extraits du message oral, le rôle du deuxième niveau du processus d'interprétation est d'assembler ces unités pour produire une interprétation formelle. Ce processus est appelé *composition sémantique* dans ce document. Il peut prendre diverses formes selon la représentation formelle du sens utilisée. Les différentes compositions produites sont valuées grâce à un ensemble de mesures de confiance intégrant à la fois les niveaux précédents de décodage et la cohérence (syntaxique et/ou sémantique) de l'interprétation produite. A l'issue de cette phase, le contexte de production des messages est intégré afin d'enrichir et éventuellement modifier l'interprétation produite, résoudre dans le cas du dialogue oral les références vers les tours précédents de dialogue et segmenter les messages par rapport à un ensemble d'actes de dialogue. C'est l'ensemble de ce processus que nous appellerons *interprétation* et qui est décrit dans ce chapitre à partir d'exemples extraits des corpus présentés au chapitre 3.

5.1 Application de relations sémantiques

5.1.1 Principe de composition

Les relations sémantiques utilisées dans cette étude permettent d'instancier des structures sémantiques en effectuant un processus d'inférence grâce à un ensemble de règles logiques. L'opération de composition sémantique est elle-même un processus d'inférence dans laquelle les objets composés sont obtenus à partir des prémisses représentées par les concepts détectés. Les relations doivent toutes avoir un support dans le graphe de mot/concept produit durant la phase de décodage conceptuel : ce support correspond à l'intersection des supports de toutes les prémisses de la relation considérée.

Le projet LUNA vise à obtenir des corpus annotés avec des relations prédicatives entre les concepts. Sur de tels corpus il sera possible d'appliquer des méthodes d'apprentissage sur corpus, mélangeant modèles logiques du premier ordre et modèles stochastiques, par exemple en suivant le paradigme des *Markov Logic Network* (Richardson et Domingos, 2006). Dans les travaux présentés dans cette habilitation aucun corpus de ce type n'était disponible, les règles définissant les relations sémantiques entre concepts sont donc toutes issues d'un processus de mise au point manuelle et sont exprimées dans le formalisme de la logique du premier ordre. Par exemple, sur le corpus *FT3000*, nous avons utilisé un ensemble de 2600 règles définies manuellement et fournies par France Télécom.

Le principe d'application de telles règles va être illustré sur un exemple très simple, reprenant le graphe de mot/concept présenté dans le chapitre précédent : à partir des FSM des figures 4.4 et 4.5, on obtient pour chaque concept I un accepteur G_{W_i} qui est l'union de tous les chemins T_{WC} qui produisent la chaîne de concept I_i : $[G_{W_i}] = [T_{WC} \circ S_i]$

Ce processus est illustré par la figure 5.1. En utilisant la notation proposée par (Jackendoff, 1990), le concept `path` peut être inféré par la règle suivante :

$$[PATH] \rightarrow \left[\begin{array}{l} \left\{ \begin{array}{c} TO \\ FROM \\ NEAR \\ TOWARD \\ \ldots \end{array} \right\} \left(\left[\left\{ \begin{array}{c} THING \\ LOC \end{array} \right\} \right] \right) \\ path \end{array} \right]$$

La règle précise que la composition du modifieur `NEAR` avec une instance de lieu `LOC` peut produire une instance de `PATH`. Si l'hypothèse `NEAR` est représentée par l'accepteur $G_{W_{NEAR}}$ et `<LOC>` par l'accepteur $G_{W_{LOC}}$ (obtenue selon la méthode présentée dans le chapitre précédent), alors l'hypothèse $G_{W_{PATH}}$ correspondant à une instance de `<PATH>` est générée si et seulement si :

$$G_{W_{PATH}} = [G_{W_{NEAR}} \cap G_{W_{LOC}}] \neq \emptyset$$

Si l'intersection des supports des prémisses de l'inférence n'est pas vide, la nouvelle

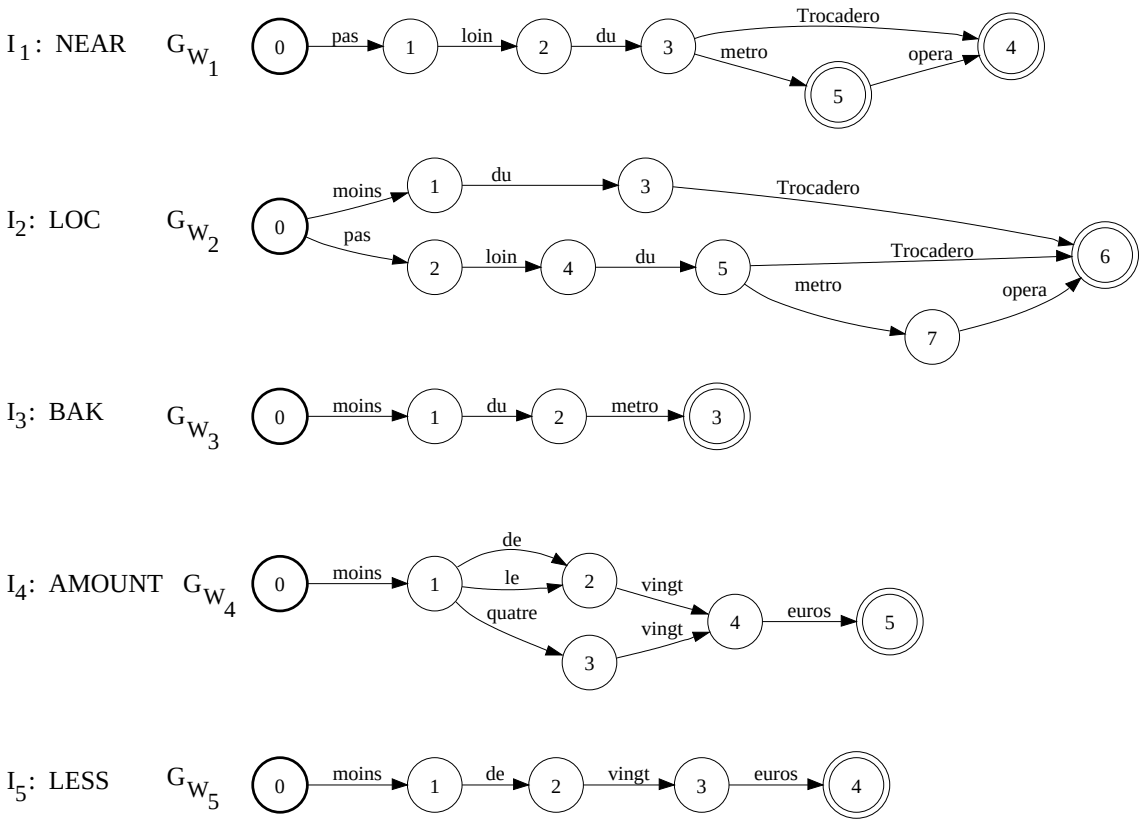


FIG. 5.1: *liste de n-meilleures séquences de concepts I_i avec leur accepteur correspondant G_{W_i}, à partir de T_{WC}*

structure sémantique est ajoutée à l'ensemble des interprétations possibles du message et le processus est répété jusqu'à ce qu'aucune nouvelle relation ne s'applique.

Par exemple, pour les cinq concepts de la figure 5.1, deux relations sémantiques peuvent s'appliquer :

- une pour l'instance <LOC>, comme présenté précédemment, fondée sur les concepts NEAR (I_1) et LOC (I_2) ;
- une pour l'instance <MONEY> : $\lfloor_{money} LESS(\lfloor_{thing} AMOUNT \rfloor) \rfloor$ fondée sur les concepts LESS (I_5) et AMOUNT (I_4).

Ceci conduit à effectuer six opérations sur les accepteurs G_{W_i} :

$$
\begin{aligned}
G_{W_{I_1 \cap I_2}} &= [G_{W_1} \cap G_{W_2}] && \rightarrow NEAR + LOC \\
G_{W_{I_1 - I_2}} &= [G_{W_1} - G_{W_2}] && \rightarrow NEAR \\
G_{W_{I_2 - I_1}} &= [G_{W_2} - G_{W_1}] && \rightarrow LOC \\
G_{W_{I_4 \cap I_5}} &= [G_{W_4} \cap G_{W_5}] && \rightarrow AMOUNT + LESS \\
G_{W_{I_4 - I_5}} &= [G_{W_4} - G_{W_5}] && \rightarrow AMOUNT \\
G_{W_{I_5 - I_4}} &= [G_{W_5} - G_{W_4}] && \rightarrow LESS
\end{aligned}
$$

Comme $G_{W_{I_5 - I_4}} = \emptyset$ seules six interprétations sont gardées : les cinq que l'on vient d'obtenir et l'interprétation vide I_3 (BAK).

Sur le FSM de la figure 4.3, on obtient :

$P(I_1 \cap I_2)$	$= 0.58$	$P(I_1 - I_2)$	$= 0.21$
$P(I_3)$	$= 0.11$	$P(I_4 - I_5)$	$= 0.07$
$P(I_4 \cap I_5)$	$= 0.028$	$P(I_2 - I_1)$	$= 0.002$

La figure 5.2 montre les six interprétations conservées avec les FSM correspondant.

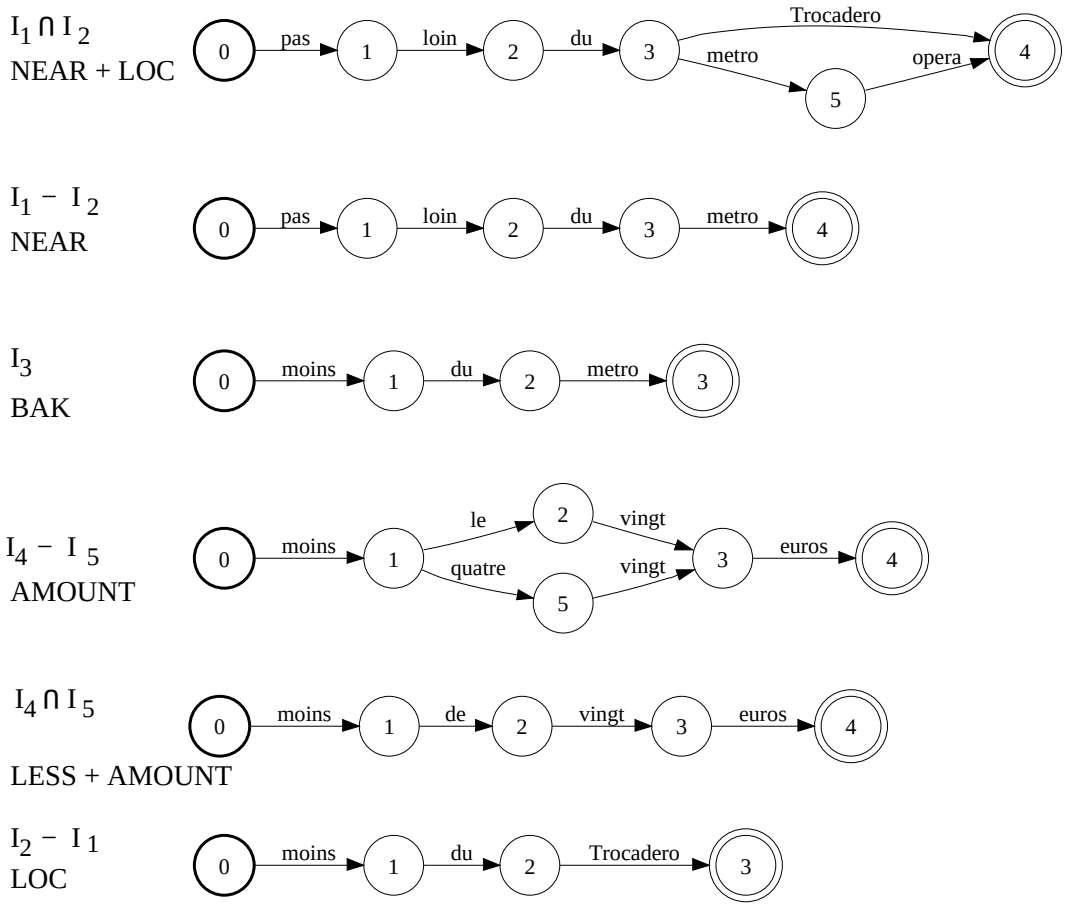


FIG. 5.2: *Liste d'interprétations conceptuelles après application de relations sémantiques aux concepts de la figure 5.1*

5.1.2 Application à des bases de règles de composition : exemple sur le corpus *FT3000*

Les relations sémantiques permettant de composer des concepts peuvent aussi être exprimées sous la forme d'automates à états finis. Dans ce cas le principe de composition consiste simplement à effectuer une intersection entre le FSM mot/concept produit lors de la phase de décodage conceptuel et le FSM encodant les relations sémantiques.

Par exemple, sur le corpus *FT3000*, un ensemble d'environ 2600 règles définies manuellement représentent les règles d'interprétation portant sur les concepts obtenus

dans la première phase. Chaque règle contient les opérateurs suivants :

- le *ou* logique représenté par l'opérateur | ;
- le *et* logique représenté par l'opérateur # ;
- le *et* séquentiel représenté par l'opérateur *&*

Voici un exemple de règles produisant la structure prédicative :

Gest(Resilier,Ambi(AtoutsPlus,HeureLocale,ForfaitLocal))

```
((Resilier|Annuler|Supprimer|Arreter|Plu)
# ((Appel|Appelle|Telephone|Telephoner) & Frequent & Domicile))
=> {Gest(Resilier,Ambi(AtoutsPlus,HeureLocale,ForfaitLocal))}
```

En plus de l'inférence produite, chaque règle est également associée à un numéro indiquant sa priorité dans la base de règles. Cet indice de priorité permet de choisir parmi des interprétations produites à partir de la même séquence de concepts. Le FSM obtenu à partir de la règle donnée en exemple est présenté dans la figure 5.3. Le modèle développé sur le corpus *FT3000* contient l'union de tous les FSM encodant ces règles.

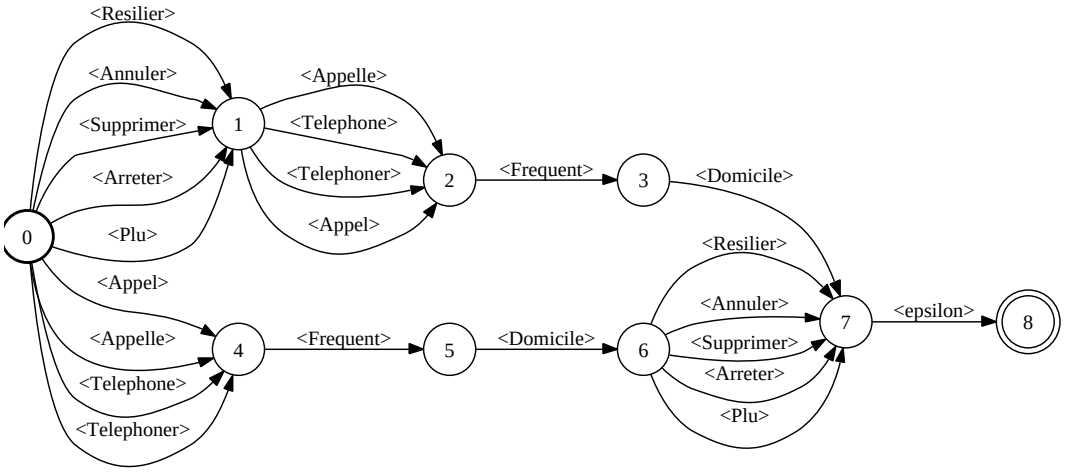


FIG. 5.3: *Exemple de règle d'interprétation sur le corpus FT3000 représenté sous forme de FSM*

5.1.3 Liste structurée de *n*-meilleures interprétations

L'application de règles de composition directement dans le graphe mot/concept permet d'obtenir une liste de *n*-meilleures hypothèses d'interprétations *structurées* par rapport aux différents concepts attribut/valeur que l'on trouve dans le graphe. A la différence des listes de *n*-meilleures hypothèses produites à partir des graphes de mots seuls, chaque hypothèse ici correspond à une interprétation possible du message traité. Cette liste est dite *structurée* car elle contient trois niveaux :

1. le premier niveau contient les séquences d'attributs de concepts ou de prédicats obtenues à l'issue de la phase de composition sémantique :

2. dans le deuxième niveau les meilleures valeurs trouvées dans le graphe, par le processus d'extraction de valeurs présenté dans le paragraphe 4.4, sont données pour chaque attribut ;
3. enfin le troisième niveau contient la meilleure séquence de mots support pour chaque couple attribut (ou prédicat)/valeur.

Un exemple d'une telle liste structurée est donné dans la table 5.1, toujours sur le même exemple.

rank	interpretation/value	score
I_1	$\lfloor_{path} NEAR(\lfloor_{place} IN(\lfloor_{thing} LOC\rfloor)\rfloor)\rfloor$	0.58
$I_{1.1}$	$LOC(type : subway, value = opera)$	0.57
W	`pas loin du metro opera` (`near metro opera`)	
$I_{1.2}$	$LOC(type : square, value = Trocadero)$	0.01
W	`pas loin du Trocadero` (`near the Trocadero`)	
I_2	$NEAR$	0.21
W	`pas loin du metro` (`near metro`)	
I_3	BAK	0.11
W	`moins du metro` (`less metro`)	
I_4	$\lfloor_{thing} AMOUNT\rfloor$	0.07
$I_{4.1}$	$AMOUNT(type : euros, value = 80)$	0.065
W	`moins quatre vingt euros` (`less eighty euros`)	
$I_{4.2}$	$AMOUNT(type : euros, value = 20)$	0.005
W	`moins le vingt euros` (`less the twenty euros`)	
I_5	$\lfloor_{money} LESS(\lfloor_{thing} AMOUNT\rfloor)\rfloor$	0.028
$I_{5.1}$	$AMOUNT(type : euros, value = 20)$	0.028
W	`moins de vingt euros` (`less than twenty euros`)	
I_6	$\lfloor_{thing} LOC\rfloor$	0.002
$I_{6.1}$	$LOC(type : square, value = Trocadero)$	0.002
W	`moins du Trocadero` (`less the Trocadero`)	

TAB. 5.1: *Exemple de liste structurée de* ***n****-meilleures hypothèses obtenue sur le graphe de mots de la figure 4.3*

Ce type de structure peut être vu comme l'ensemble des interprétations potentielles d'un message. L'avantage principal de ce type de liste de **n**-meilleures hypothèses, comparé à une liste standard obtenue sur les mots seuls, est que chaque hypothèse a un sens différent. Le contexte de production du message peut alors être utilisé pour filtrer parmi cette liste les hypothèses incohérentes avec ce contexte. A l'inverse, si le contexte permet de prédire un certain nombre d'interprétations probables, ces interprétations peuvent être recherchées directement dans la liste d'hypothèses structurées, et pour chacune d'elles les meilleurs valeurs possibles sont extraites.

5.2 Intégration du contexte de production

Le contexte de production des messages permet d'enrichir et éventuellement de modifier les interprétations produites. Dans le cas du dialogue oral, le contexte est in-

dispensable pour résoudre ces références effectuées envers les tours précédents de dialogue et segmenter les messages par rapport à un ensemble d'actes de dialogue. Nous allons présenter deux modèles permettant cette interprétation.

Le premier modèle intègre le contexte du dialogue directement dans la recherche de la meilleure interprétation, cette étude publiée dans (Damnati et al., 2007a), a été faite sur le corpus *FT3000*.

Le deuxième modèle enrichit une liste de **n-**meilleures interprétations telles que celles présentées dans le paragraphe précédent avec des étiquettes spécifiant le sens d'un concept relativement au contexte de dialogue. Ces *spécifieurs* permettent aussi de proposer un algorithme simple s'attaquant au problème des références entre différents tours de dialogue.

5.2.1 Décodage intégré mot/concept avec historique de dialogue

Considérons un modèle de dialogue représenté par un automate à états finis dans lequel une transition entre un état S_i et un état S_j est provoquée par l'interprétation d'un message M. Dans chaque état le système de dialogue génère un message pour l'utilisateur. Pendant que le dialogue progresse, la stratégie de dialogue avance selon les états de l'automate.
Soit $S = \{S_0, S_1, \ldots, S_k\}$ une telle séquence. L'état S_k est atteint, au tour de dialogue k, après traitement du message M_k représenté par la séquence d'observations acoustiques Y_k ayant été interprétée comme la structure prédicative Γ_k.

Soit $Y = \{Y_1, Y_2, \ldots, Y_k\}$ la séquence des observations acoustiques interprétées comme $\Gamma = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_k\}$. Nous voulons calculer la probabilité $P(S|Y)$ pour l'utiliser dans une stratégie de dialogue permettant de gérer des états multiples de dialogue, raffinés au fur et à mesure de l'avancée du dialogue et de la spécification du contexte.

Une stratégie qui tient compte d'hypothèses d'états de dialogue multiples a été récemment proposée par (Williams et Young, 2007), fondée sur un procédé de décision de Markov partiellement observable (POMDP). Le modèle proposé ici est plus simple car tous les états sont définis dans l'automate de dialogue à états finis.

L'estimation de cette probabilité $P(S|Y)$ est définie de manière récursive comme suit :

$$P(S|Y) = \sum_{\Gamma} P(S\Gamma|Y) = \sum_{\Gamma} P(S_k\Gamma_k|H_kY)P(H_k|Y) \tag{5.1}$$

avec $H_k = \{S_{1,k-1}, \Gamma_{1,k-1}\}$ et

$$\begin{aligned} P(S_k\Gamma_k|H_kY) &= P(S_k|\Gamma_kH_kY)P(\Gamma_k|H_kY) \\ &\approx P(S_k|\Gamma_kH_k)P(\Gamma_k|Y_k) \end{aligned} \tag{5.2}$$

Si aucun historique n'est pris en compte, alors $P(S_k|\Gamma_kH_k)$ est ramené à $P(S_k|\Gamma_k)$. Quand un corpus d'apprentissage est disponible, cet historique peut être approximé par un

modèle **n**-grammes sur les états du dialogue. Dans cette étude nous utilisons un modèle bigramme. Ainsi, nous avons :

$$P(S_k|\Gamma_k H_k) \approx P(S_k|\Gamma_k S_{k-1}) \tag{5.3}$$

L'estimation de $P(\Gamma_k|Y_k)$ est faite de la manière suivante : l'interprétation Γ_k est obtenue par composition sémantique sur la séquence de concepts C_k extraites de Y_k. Plus précisément cette séquence C_k est obtenue à partir de la chaîne de mots W_k reconnus à partir de Y_k. Ainsi nous avons :

$$\begin{aligned} P(\Gamma_k|Y_k) &= \sum_{C_k,W_k} P(\Gamma_k C_k W_k|Y_k) \\ &\approx \sum_{C_k,W_k} P(\Gamma_k|C_k)P(C_k|W_k)P(W_k|Y_k) \end{aligned} \tag{5.4}$$

$P(\Gamma_k|C_k)$ est donnée par le module de composition sémantique. Si ce module est fondée uniquement sur des règles logiques du premier ordre, comme dans les exemples donnés dans ce chapitre, cette probabilité est soit 1 soit 0 selon que la règle qui produit Γ_k s'applique ou pas sur C_k.
Les probabilités $P(C_k|W_k)P(W_k|Y_k)$ sont estimées par le modèle de décodage mot/concept présenté dans le chapitre 4, équation 4.4.

Dans ce cadre trois stratégies peuvent être construites :

1. La première stratégie est purement séquentielle et correspond à la plupart des systèmes de compréhension : la meilleure chaîne de mots $\hat{W}$ est obtenue par $\hat{W} = \underset{W}{argmax} P(W|Y)$. Alors, la meilleure séquence de concepts est obtenue par : $\hat{C} = \underset{C}{argmax} P(C|\hat{W})$. Enfin l'ensemble des règles d'interprétation est appliquée à $\hat{C}$ de manière à obtenir Γ. Aucun historique de dialogue n'est pris en compte, ainsi l'équation 5.2 devient :

$$P(S_k\Gamma_k|H_kY) \approx P(S_k|\Gamma_k)P(\Gamma_k|\hat{C}_k)P(\hat{C}_k|\hat{W}_k)P(\hat{W}_k|Y_k)$$

2. La deuxième stratégie correspond à celle décrite dans le chapitre précédent, on cherche en même temps la meilleure séquence de mots et celle de concepts : En ne prenant pas en compte le contexte :

$$\begin{aligned} P(S_k\Gamma_k|H_kY) \approx & \quad P(S_k|\Gamma_k) \times \\ & \max_{W_k,C_k} P(\Gamma_k|C_k)P(C_k|W_k)P(W_k|Y_k) \end{aligned} \tag{5.5}$$

3. La dernière strategie intègre l'historique du dialogue :

$$\begin{aligned} P(S_k\Gamma_k|H_kY) \approx & \quad P(S_k|\Gamma_k S_{k-1}) \times \\ & \max_{W_k,C_k} P(\Gamma_k|C_k)P(C_k|W_k)P(W_k|Y_k) \end{aligned} \tag{5.6}$$

A chaque tour de dialogue k, la meilleure séquence d'états, depuis le début du dialogue, est estimée avec $P(S|Y)$. Les meilleurs états du dialogue (au sens des mesures de confiance) sont envoyés au gestionnaire de dialogue.

5.2.2 Spécification du sens en contexte dans le corpus MEDIA

Les étapes de spécification du sens et de résolution des références sont appliquées aux listes de **n-**meilleures interprétations obtenues dans la première phase. La spécification du sens est vue comme une tâche d'étiquetage pouvant être traitée grâce à des modèles probabilistes. La résolution des références est faite par un certain nombre d'heuristiques décrivant tous les rattachements possibles pouvant être faits entre la liste de **n-**meilleures interprétations et l'historique du dialogue. Ces modèles ont été développés lors de ma participation à la tâche d'évaluation de la compréhension en contexte de la campagne MEDIA dont les résultats sont présentés dans (Denis et al., 2007).

Etiquetage des attributs de concepts avec des spécifieurs de sens

Les séquences produites à partir du graphe de concepts dans la phase de décodage conceptuel ne contiennent aucun spécifieur de sens prenant en compte le contexte du dialogue. Ces spécifieurs sont attribués par un étiqueteur probabiliste fondé sur les Champs Conditionnels Aléatoires (ou *Conditional Random Fields* CRF). Les CRF (Lafferty et al., 2001) ont été utilisés avec succès dans de nombreuses tâches d'étiquetage telles que l'étiquetage morphosyntaxique ou la détection d'entités nommées. L'avantage principal des CRF par rapport à des modèles génératifs tels que les Hidden Markov Model (HMM) est la possibilité d'utiliser l'ensemble des observations d'une séquence pour prédire une étiquette. Ce n'est donc pas le seul historique immédiat qui contraint l'attribution d'une étiquette à une observation mais potentiellement toutes les observations précédentes et suivantes. Cela est particulièrement intéressant pour l'étiquetage des spécifieurs dans la mesure où la spécification du sens d'un concept peut se faire avec des éléments situés avant ou après le concept dans le message, ou dans les tours de dialogue précédents. Les liens référentiels sont étiquetés dans cette phase par rapport au type de l'objet référencé et au type de la référence. Cette information sera utilisée dans le processus de résolution des références.

Le corpus d'apprentissage des CRF est obtenu à partir des corpus MEDIA. Chaque dialogue constitue une séquence où les observations sont les concepts marqués dans la référence et les étiquettes sont soit les spécifieurs attribués aux concepts ; soit le type du ou des objets référencés pour les liens référentiels, ainsi que le type du lien ; soit le symbole *NULL* si un concept n'a ni spécifieur ni lien référentiel. Lors du traitement d'un message, chaque chaîne de concepts produite par le décodeur mot/concept est traitée par l'étiqueteur CRF et la description des concepts est enrichie avec les étiquettes attribuées. L'étiqueteur développé utilise l'outil **CRF++** [1].

Résolution des références

A la suite de la phase précédente les concepts lien référentiels sont étiquetés avec les trois informations suivantes :

[1] téléchargeable à `http ://crfpp.sourceforge.net/`

1. Type d'objet pointé : *chambre*, *hôtel*, *réservation* (ou une combinaison de ces valeurs).
2. Type de lien référentiel : *ambigu*, *exclusion*, *inclusion*.
3. Nombre : *singulier* ou *pluriel*.

La résolution des références s'effectue alors selon l'algorithme suivant : tous les concepts situés dans l'historique du dialogue (limité aux n énoncés précédents) ayant une étiquette *spécifieur* similaire au type d'objet pointé par le lien référentiel sont associés à ce lien.

Cette association permet de remplir des cadres sémantiques pour chaque type d'objet. Nous avons les trois cadres suivants :

```
Hôtel { ville, marque, nom, service[MAX_SERVICE]; }
Chambre { type, montant, monnaie; }
Réservation {début,fin,unité,mois,date,axetps,nbnuit,nbenfant,nbcouple;}
```

Chaque objet est caractérisé par un certain nombre de traits (par exemple la ville, la marque, le nom ou les services associés à un hôtel). L'algorithme d'association fait pointer le lien référentiel vers tous les concepts représentant ces traits. Lorsque tous les traits sont identifiés, l'algorithme s'arrête s'il s'agit d'un lien référence singulier. Sinon un nouvel objet est créé et le processus se poursuit. Aucun contrôle n'est effectué sur le nombre d'objets désignés. Le but ici est de maximiser les mesures de rappel sur les références pour proposer au module de décision (analyseur sémantique, gestionnaire de dialogue) le plus d'associations possibles, chacune avec un score donné par les différents modèles utilisés lors du décodage.

Chapitre 6

Exemples d'applications et d'évaluation des modèles proposés

Sommaire

Nous allons présenter dans ce chapitre quelques applications des modèles présentés dans ce document, sur les corpus décrits au chapitre 3. Ces présentations s'appuient sur des publications dans des revues ou des conférences, le but ici n'est pas de détailler chaque publication, mais plutôt d'en faire ressortir les points saillants étayant les modèles proposés et de discuter les résultats obtenus.

Quatre points vont être abordés :

1. tout d'abord l'une des caractéristiques principales des modèles proposés, à savoir l'intégration des processus de RAP et de compréhension ;

2. nous verrons ensuite l'apport de la prise en compte du contexte dans les performances de plusieurs systèmes de compréhension pour des applications de dialogue et de recherche d'information ;
3. la mise au point de stratégies faisant intervenir des mesures de confiance et permettant d'adapter le processus de compréhension au message traité sera présentée pour deux applications de dialogue ;
4. enfin les particularités de la parole très spontanée, telle que l'on peut en trouver dans les corpus écologiques ou collectés auprès de vrais utilisateurs seront discutées pour une application de dialogue et une application de recherche d'information.

6.1 Intégration des processus de RAP et de compréhension

6.1.1 Illustration du modèle par un exemple

Le modèle proposant de projeter un graphe de mots vers un graphe de concepts, en utilisant le formalisme des automates à états finis, a été publié en 2006 dans un article de la revue *Speech Communication* (Raymond et al., 2006). Cet article propose aussi la représentation des sorties du module de compréhension sous la forme d'une liste *structurée* d'interprétations, comme nous l'avons montré dans le chapitre 4.

Ce processus est illustré par les figures suivantes, sur un exemple extrait du corpus de France Télécom PLANRESTO :

- le message prononcé est : « *dans le quartier des Halles euh euh un restaurant euh autour de dix euros par personne* » ;
- le graphe de mots produit par le module de RAP à partir du message vocal est présenté dans la figure 6.1 ;
- à l'issue du processus de projection du graphe de mots vers un graphe de concepts, on obtient le graphe présenté dans la figure 6.2, d'où est extraite la liste structurée de n-meilleures hypothèses d'interprétation (limitée ici aux deux meilleures). Le chiffre entre crochets après chaque interprétation correspond à la place, dans la liste de n-meilleures hypothèses en mots, de la séquence de mots supports à l'interprétation.

Comme nous pouvons le voir une réduction très importante de l'espace de recherche est faite lorsqu'on limite celui-ci aux hypothèses ayant un *sens* différent par rapport à l'application visée. Par exemple, les 28 premières hypothèses en mots, obtenues à partir du graphe de la figure 6.1, sont données dans la table 6.1. On peut voir que celles-ci diffèrent principalement à cause de mots outils de petite taille pouvant être insérés dans les zones instables du décodage correspondant le plus souvent à la réalisation de disfluences de la part du locuteur. L'hypothèse sémantique correcte est classée en trente neuvième position dans cette liste, alors qu'elle se trouve dans la deuxième hypothèse d'interprétation de la liste structurée.

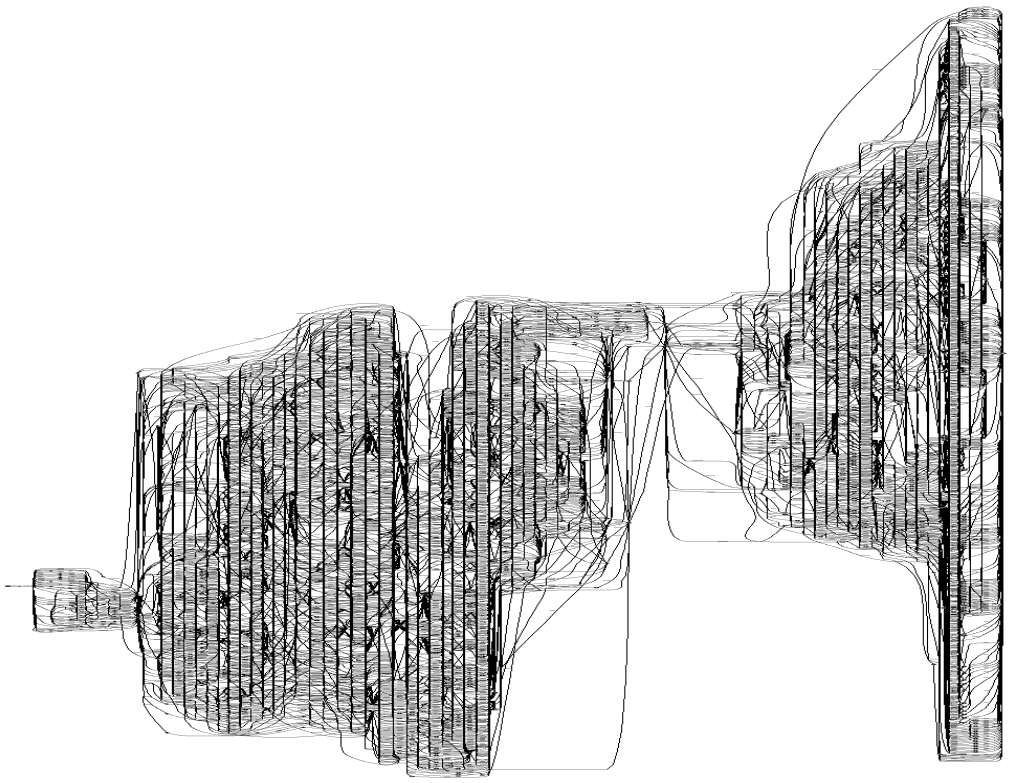


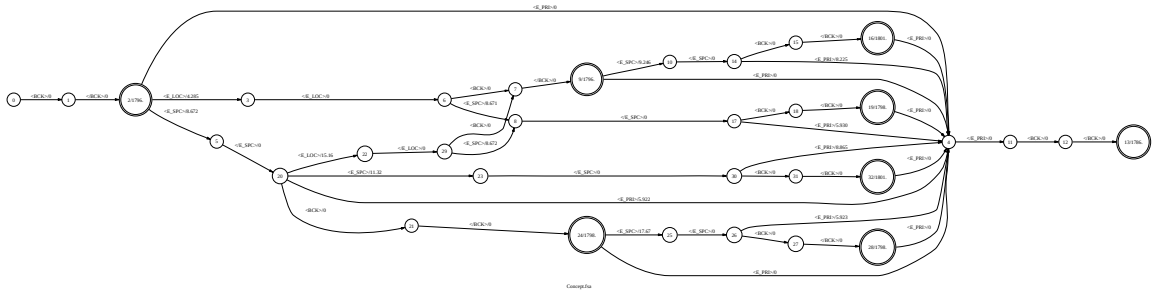
FIG. *6.1:* *Graphe de mots, sous forme de FSM, obtenu à partir d'un message du corpus PLANRESTO*

6.1.2 Evaluation

Comme précisé dans le chapitre 4, l'approche séquentielle consiste à produire d'abord la meilleure chaîne de mots $\hat{W}$ à partir des modèles de RAP seuls, puis à analyser le texte produit pour en extraire l'interprétation $\hat{\Gamma}$, à l'inverse de l'approche intégrée qui cherche simultanément $\hat{W}$ et $\hat{\Gamma}$. Cette comparaison est faite par deux mesures : le taux d'erreurs d'interprétation et le taux *Oracle* d'erreurs pour une liste d'hypothèses comme présenté dans la section 2.4.1.

Expériences sur le corpus PLANRESTO

Les premiers résultats, publiés dans (Raymond et al., 2006), montrent l'intérêt de produire une liste d'hypothèses structurée selon les diverses interpétations portées par chaque hypothèse. Dans ces résultats, résumés par la figure 6.3, il n'y a pas de modèle de langage sur les concepts. Seuls les scores acoustiques et ceux du modèle de langage général sont pris en compte. C'est pour cela que la meilleure hypothèse de l'approche intégrée est identique à celle de l'approche séquentielle. Le taux d'erreurs mesuré est le taux d'erreurs sur les concepts ou *Concept Error Rate* (CER). Le taux d'erreurs sur



Interprétation 1 : $< PRI >$

- dans le quartier des vins euh le restaurant autour de *dix euros* par personne [1]
- dans le quartier des vins euh le restaurant entre *deux mille euros* par personne [169]

Interprétation 2 : $< LOC >< PRI >$

- dans le quartier des *Halles* restaurant autour de *dix euros* par personne [39]

FIG. 6.2: *Graphe de concepts et liste d'hypothèses d'interprétation structurée obtenus sur le graphe de mots de la figure 6.1*

les mots de la meilleure hypothèse de transcription est de 26%, le modèle sémantique utilisé contient 15 concepts distincts.

Le score Oracle montre clairement l'intérêt de la méthode intégrée : en ne gardant que le 5 meilleures hypothèses de la liste structurée, on obtient un taux d'erreurs proche de la valeur minimale que l'on peut atteindre en gardant toutes les hypothèses alors qu'il faut garder 35 hypothèses de la liste non-structurée pour atteindre le même taux.

Expériences sur le corpus MEDIA

La figure 6.4 montre l'architecture du système développé pour le corpus MEDIA et présenté dans (Servan et al., 2006). Sur ce corpus, comme présenté dans le chapitre 3, le modèle sémantique est riche de 80 concepts. Un modèle de décodage estimant la probabilité jointe de la meilleure chaîne de mots et de concepts est entraîné sur le corpus d'apprentissage annoté manuellement et est utilisé pour réévaluer tous les chemins du transducteur mot/concept.

La figure 6.5 montre la corrélation entre le taux d'erreur sur les mots (*Word Error Rate* - WER) et le taux d'erreur sur les concepts (CER). Rappelons que le CER intègre les erreurs des attributs et des valeurs associés aux concepts. Cette figure compare également l'approche intégrée et l'approche séquentielle. Pour l'approche séquentielle les taux d'erreurs mots indiqués dans la figure correspondent aux taux d'erreur de la transcription automatique utilisée en entrée du décodage conceptuel. Pour l'approche inté-

```
01 dans le quartier des vins euh le restaurant autour de dix euros par personne
02 dans le quartier des vins euh le restaurant entre deux dix euros par personne
03 dans le quartier des repas le restaurant autour de dix euros par personne
04 dans le quartier des repas le restaurant entre deux dix euros par personne
05 dans le quartier des un restaurant autour de dix euros par personne
06 dans le quartier des vins euh le restaurant autour de dix euros par personne plus
07 dans le quartier des vins euh le restaurant autour de dix euros par personne plus d'
08 dans le quartier des vins euh le restaurant autour de dix euros par personne ce
09 dans le quartier des vins euh le restaurant autour de dix euros par personne ce n'
10 dans le quartier des un restaurant entre deux dix euros par personne
11 dans le quartier des vins euh le restaurant autour de dix euros par personne je c' est
12 dans le quartier des vins euh restaurant autour de dix euros par personne
13 dans le quartier des vins euh le restaurant entre deux dix euros par personne plus
14 dans le quartier des vins euh le restaurant entre deux dix euros par personne plus d'
15 dans le quartier des vins euh le restaurant entre deux dix euros par personne ce
16 dans le quartier des vins euh le restaurant entre deux dix euros par personne ce n'
17 dans le quartier des vins euh le restaurant pour deux dix euros par personne
18 dans le quartier des vins euh le restaurant entre deux dix euros par personne je c' est
19 dans le quartier des vins euh restaurant entre deux dix euros par personne
20 dans le quartier des vins euh le restaurant autour de dix euros par personne je c' est
21 dans le quartier des vins euh restaurant autour de dix euros par personne
22 dans le quartier des vins euh le restaurant entre deux dix euros par personne plus
23 dans le quartier des vins euh le restaurant entre deux dix euros par personne plus d'
24 dans le quartier des vins euh le restaurant entre deux dix euros par personne ce
25 dans le quartier des vins euh le restaurant entre deux dix euros par personne ce n'
26 dans le quartier des vins euh le restaurant pour deux dix euros par personne
27 dans le quartier des vins euh le restaurant entre deux dix euros par personne je c' est
28 dans le quartier des vins euh restaurant entre deux dix euros par personne
```

TAB. 6.1: *Exemple de liste de n-meilleures hypothèses sur les mots obtenue sur le graphe de mots de la figure 6.1*

grée, ces taux d'erreurs sont ceux des meilleurs chemins des graphes de mots utilisés en entrée. Cette courbe montre clairement la corrélation entre taux d'erreurs mots et taux d'erreurs sur les concepts. On voit aussi que l'approche intégrée apporte un gain modeste mais constant par rapport à l'approche séquentielle.

Une autre courbe intéressante est celle publiée dans (Servan et Bechet, 2006), toujours sur le corpus MEDIA, et montrant les performances du décodage conceptuel en fonction de la taille des données d'apprentissage. Le nombre de dialogues d'apprentissage est donné en fonction du taux d'erreurs sur les concepts (CER). Le CER est ici estimé sur les transcriptions exactes du corpus MEDIA. Cette courbe est présentée dans la figure 6.6. Comme on peut le voir la courbe se stabilise autour de 400 dialogues d'apprentissage. L'utilisation de connaissances *a priori*, représentées ici par les grammaires manuelles ajoutées aux grammaires extraites du corpus d'apprentissage, permet d'améliorer les résultats quand la quantité de corpus d'apprentissage est restreinte. Ce gain s'amenuise à mesure qu'augmente la taille du corpus.

Expériences sur le corpus FT3000

L'approche intégrée a aussi été évaluée sur le corpus FT3000 en rajoutant l'étape de composition sémantique, comme présenté dans le chapitre 5. Ces travaux on été publiés dans (Damnati et al., 2007a,b).

Si les concepts sont peu ambigus dans le FT3000, en revanche ils sont très nombreux (plus de 400), et le nombre de structures prédicatives différentes que l'on peut produire par composition de concepts est aussi très élevé (environ 1800). De plus contrairement à PLANRESTO ou MEDIA, le corpus FT3000 contient de la parole fortement spontanée.

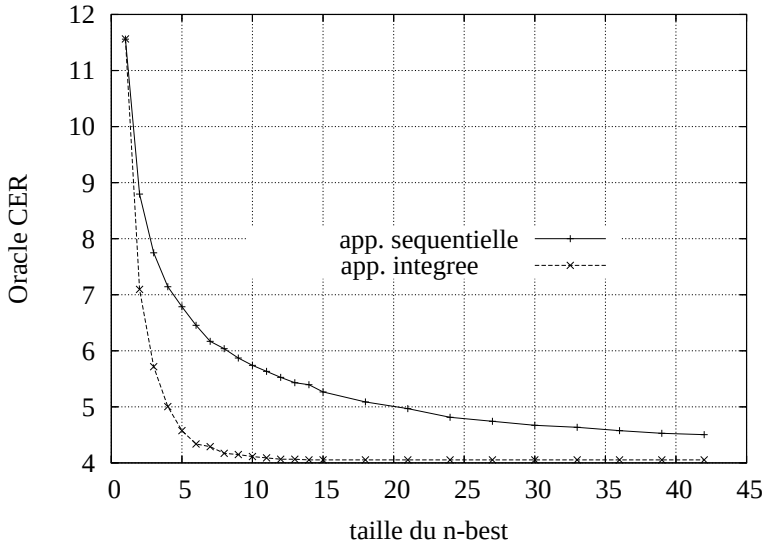


FIG. 6.3: *Comparaison des taux d'erreurs Oracle pour les approches séquentielles et intégrées sur le corpus PLANRESTO*

taux d'erreurs	*WER*	*CER*	*IER*
séquentielle	40.1	24.4	15.0
intégrée	38.2	22.5	14.5

TAB. 6.2: *Word Error Rate (WER), Concept Error Rate (CER) et Interpretation Error Rate (IER) selon la stratégie séquentielle ou intégrée*

La comparaison entre les deux approches (séquentielle et intégrée) sur le FT3000 est donnée dans la table 6.2. En plus des taux d'erreurs sur les mots (WER) et les concepts (CER), le taux d'erreurs sur les structures prédicatives produites est donné. Il est appelé ici le taux d'erreurs sur les interprétations ou *Interpretation Error Rate* (IER). Tous les éléments de la structure produite doivent être corrects pour qu'une hypothèse soit considérée comme correcte. Le gain global sur les interprétations est modeste. Cependant l'approche intégrée permet de produire grâce à la liste d'hypothèses structurées un sous-ensemble d'hypothèses très réduit avec un excellent taux d'erreur Oracle, comme présenté dans la figure 6.7

Le contexte du dialogue peut permettre de filtrer une liste d'hypothèses par rapport à ce qui est attendu, il est donc très intéressant du point de vue du gestionnaire du dialogue d'obtenir une liste d'hypothèses provenant du module de compréhension.

6.1.3 Discussion

Les expériences menées sur ces différents corpus ont toutes mis en évidence l'efficacité de l'approche de décodage intégrée pour produire une liste réduite d'hypothèses *pertinentes* par rapport à la tâche. De manière consistante l'approche intégrée produit

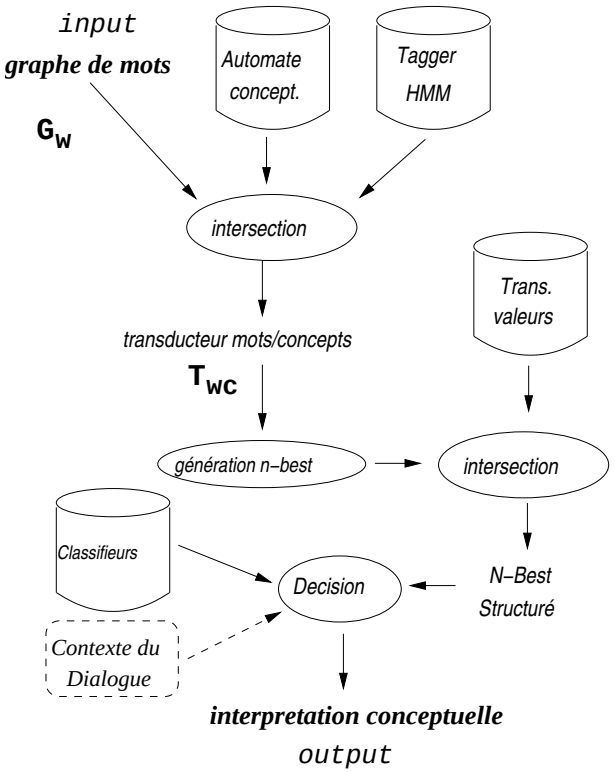


FIG. 6.4: *Architecture du système de compréhension sur le corpus MEDIA*

de bien meilleurs scores Oracle que l'approche séquentielle. Cependant le gain en performance obtenu en ne prenant que la meilleure hypothèse est modeste. A mon sens l'intérêt principal de la méthode proposée est la possibilité de garder un espace de recherche ouvert sur les interprétations possibles d'un message oral. Le module utilisant les hypothèses de compréhension (gestionnaire de dialogue, système de recherche d'information) va ainsi pouvoir interroger cet espace, structuré par rapport à tous les attributs et toutes les valeurs qu'il est possible de reconnaître, et utiliser des informations contextuelles pour valider ou supprimer un certain nombre d'entre elles.

6.2 Prise en compte du contexte de production des messages

Les modèles numériques utilisés à la fois pour la RAP et la compréhension de parole posent comme hypothèse fondamentale la similarité entre les données utilisées pour leur apprentissage et celles qu'ils vont traiter en phase d'exploitation. Plus cette similarité est grande, meilleurs seront les résultats. Afin de garantir cette similarité, il est crucial d'identifier précisément le contexte de production du message à traiter afin de choisir le modèle qui lui est le mieux adapté. Dans un cadre de dialogue oral, cela consiste à identifier l'état du dialogue et à adapter les modèles au langage le plus probable dans cet état. Pour une tâche de recherche d'information le contexte consiste à

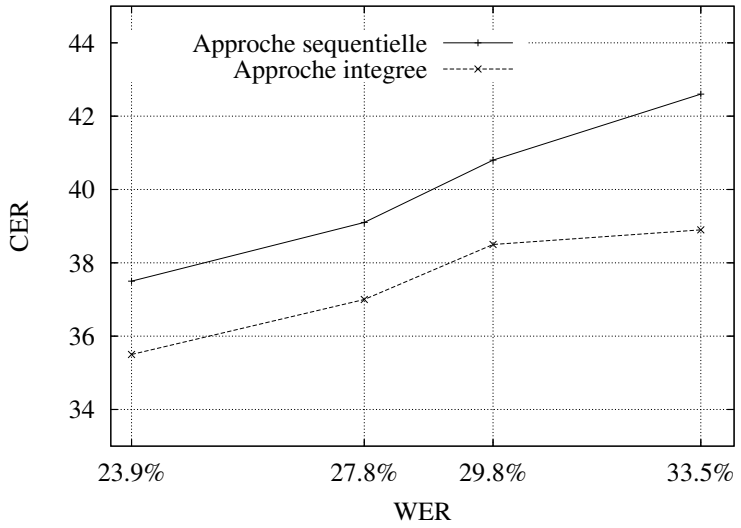


FIG. 6.5: *Architecture du système de compréhension sur le corpus MEDIA*

identifier le type de parole (parole préparée telle que la parole journalistique ou parole conversationnelle), le thème mais aussi des informations factuelles sur le media telles que, par exemple pour le traitement de parole radiodiffusée, le nom de la radio et de l'émission, la date et la tranche horaire. Toutes ces méta-données permettent de caractériser le contexte de production des messages et notamment d'adapter si possible les modèles de RAP (lexique et modèles de langage) ainsi que les modules de compréhension (modification de la liste des concepts attendus).

Pour la RAP plusieurs études (Whittaker, 2001; Federico et Bertoldi, 2001; Chen et al., 2004) ont proposé l'adaptation d'un modèle de langage général à un corpus de petite taille extrait grâce à des méta-données. On trouvera une vue d'ensemble des méthodes d'adaptation de modèles de langage dans (Bellegarda, 2004). Selon la méthode d'adaptation choisie et le type de méta-données utilisé, des gains de performance ont pu être obtenus. Cependant il ressort de toute ces études que le choix et la taille du corpus d'adaptation sont cruciaux : si le corpus d'adaptation obtenu est trop petit ou ne correspond pas exactement aux données à traiter, aucun gain n'est constaté. Par exemple, (Chen et al., 2004) présente des résultats obtenus avec plusieurs méthodes de sélection de corpus d'adaptation ; la meilleure méthode consiste à effectuer un premier décodage afin de segmenter en thèmes le document à traiter, puis à collecter un corpus d'adaptation pour chaque thème. Ce corpus est ensuite utilisé pour adapter le modèle de langage général afin de réaliser une deuxième phase de décodage.

Pour les modèles de compréhension la première adaptation consiste à adapter le lexique de reconnaissance. En effet, si les unités lexicales capables de déclencher l'identification d'un concept ne sont pas dans le lexique de reconnaissance, ce concept n'a aucune chance d'être reconnu. Cette modification éventuelle du lexique de RAP entraîne évidemment également une adaptation du modèle de langage pour intégrer ces

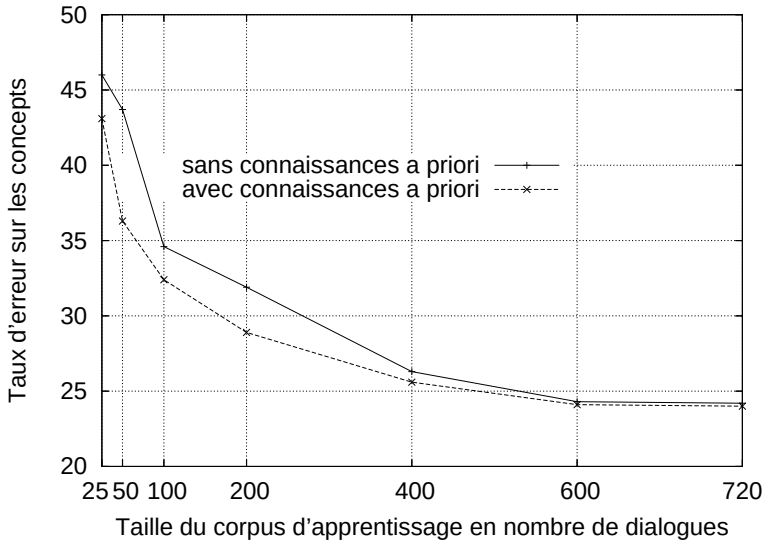


FIG. 6.6: *Corrélation entre le taux d'erreurs sur les concepts et la taille des données d'apprentissage, avec et sans grammaires manuelles (connaissances a priori)*

nouvelles entrées. Une autre adaptation concerne les modifications des probabilités *a priori* des différentes interprétations.

Nous allons décrire brièvement les travaux que nous avons menés dans ce sens sur deux applications : détection d'entités nommées dans des corpus d'émissions diffusées et dialogue oral homme-machine.

6.2.1 Intégration du contexte pour la détection d'entités nommées

Cette étude a été publiée à la conférence jointe *Human Language Technology* (HLT) et *Empirical Methods for Natural Language Processing* (EMNLP) en 2005 à Vancouver (Favre et al., 2005). La tâche visée ici est la détection d'entités nommées, évaluée en termes de précision, rappel et mesure F. Le corpus ESTER, présenté au chapitre 3, est composé d'émissions de radio, essentiellement des informations, enregistrées entre 2002 et 2004. Deux corpus d'évaluation ont été utilisés : un corpus contenant des émissions enregistrées à la même période et sur les mêmes radios que le corpus d'apprentissage ; et un corpus enregistré six mois plus tard, avec une radio supplémentaire non présente dans l'apprentissage. Les différences de performances constatées entre les deux corpus sont importantes : la mesure F perd 10 points, de 73 pour le corpus proche de l'apprentissage à 63 pour le corpus avec un décalage de 6 mois. Ce résultat illustre bien la fragilité des modèles à tout changement, même modeste, dans les conditions expérimentales. Ce résultat est similaire à ce qu'il a pu être montré dans d'autres études (Miller et al., 2000).

Afin d'essayer d'adapter les modèles d'entités nommées à ces nouvelles conditions, nous avons utilisé la méta-donnée la plus simple à obtenir : la date de diffusion de chaque émission. A partir de cette information nous avons collecté du corpus de texte

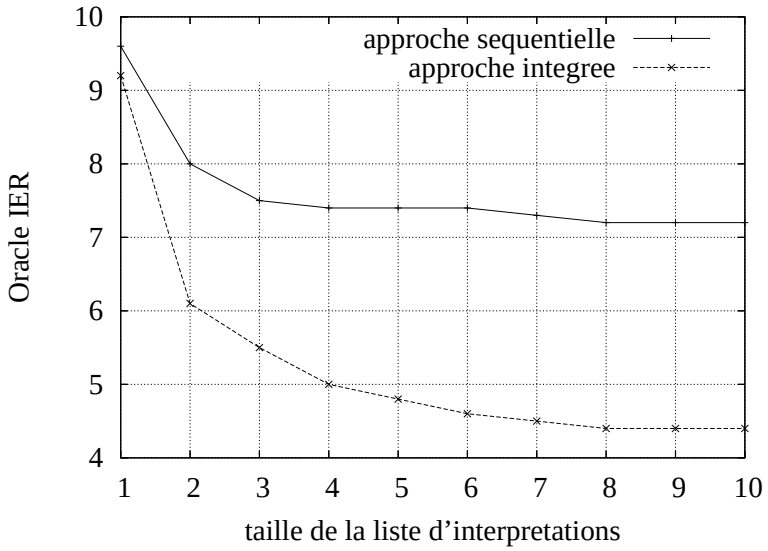


FIG. 6.7: *Taux Oracle d'erreurs d'interprétation sur le corpus FT3000 pour les deux approches, séquentielle et intégrée*

journalistique correspondant à chaque jour de cette période. En effet l'occurrence d'une entité nommée est très dépendante d'un moment donné de l'actualité. Ainsi, nous avons montré que sur ce corpus ESTER, 72% des entités nommées n'apparaissaient qu'un seul jour. En examinant, pour chaque jour, l'intersection entre les entités nommées trouvées dans le corpus de test ESTER et celles présentes dans les textes journalistiques, nous trouvons qu'en moyenne 25% d'entres elles sont communes chaque jour. Cette intersection est mise en relief par la figure 6.8 qui montre que le taux d'entités communes connaît un pic très clair sur le jour même de l'émission et du texte journalistique, ce taux diminuant rapidement avant et après ce jour cible.

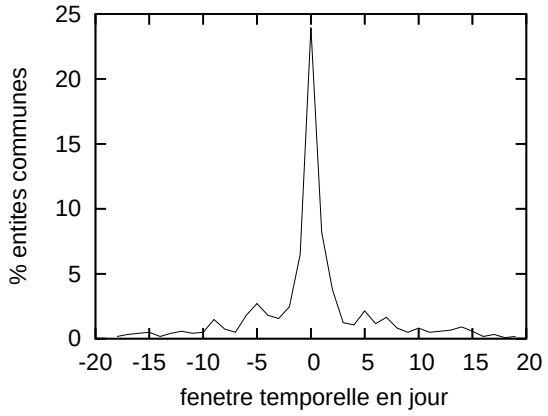


FIG. 6.8: *Pourcentage d'entités commune au corpus de test et au corpus journalistique par rapport au nombre de jour séparant ces deux corpus (pour la valeur 0, les deux corpus sont datés du même jour).*

En adaptant les modèles d'entités nommées grâce à ces corpus collectés pour chaque

jour, nous avons conduit l'expérience suivante :

- un ensemble de 352 entités a été sélectionné sur le corpus journalistique correspondant à la même période de temps que le test ; ces entités correspondent à ce qui est attendu étant donné le contexte des messages ;
- l'extraction des entités nommées à partir des graphes de mots, en utilisant l'approche intégrée, est appliquée aux graphes du test en utilisant des modèles adaptés pour chaque jour ;
- les valeurs de précision, rappel, mesure F et taux Oracle sont estimées avec et sans cette adaptation.

Les résultats sont donnés dans le tableau 6.9.

Condition	Prec.	Rappel	F-m	Oracle
sans adaptation	87.0	75.7	80.9	83.6
avec adaptation	87.5	83.9	85.7	92

FIG. 6.9: *Résultat de détection d'entités nommées sur les entités sélectionnées à partir des corpus journalistique, pour le corpus de test ESTER*

Le processus d'adaptation améliore significativement les mesures de rappel et d'Oracle dans la détection des 352 entités sélectionnées. Il faut toutefois noter que cette amélioration ne concerne que les 25% d'entités communes entre les documents audio à traiter et les méta-données collectées. L'amélioration globale de la détection de toutes les entités nommées est marginale (+1 en mesure F et +3% en rappel Oracle). Le fait de disposer d'une liste structurée de toutes les entités nommées pouvant être reconnues dans chaque document audio permet cependant de renverser le paradigme de détection : les entités cibles sont directement recherchées dans le transducteur mot/entité, chaque entité trouvée étant associée à la meilleure chaîne de mots support dans le graphe issus de la phase de RAP.

6.2.2 Adaptation et contexte de dialogue

La notion de contexte de dialogue n'est pas toujours facile à définir. On peut la représenter comme un état dans un automate de dialogue. Cette notion est cependant trop restrictive dans la mesure ou deux demandes de confirmation identiques peuvent se trouver dans des états éloignés de l'automate. Dans un article présenté à la conférence Empirical Methods for Natural Language Processing (EMNLP 2004) (Bechet et al., 2004), nous avons proposé une méthode de classification automatique permettant de déterminer des groupes d'états de dialogue partageant des caractéristiques communes et nous l'avons appliquée au corpus de dialogue d'AT&T HMIHY. Cette méthode est fondée sur une classification hiérarchique non-supervisée, les paramètres utilisés pour caractériser chaque tour de parole sont : le type et le texte du prompt joué par la machine, l'étiquette d'état du dialogue donnée par le gestionnaire de dialogue, l'historique du dialogue représenté par la séquence d'états précédents et enfin la durée du message, estimée en nombre de mots. Le critère optimisé lors de cette classification est la

perplexité en mots des transcriptions manuelles des messages.

Tous ces paramètres peuvent être obtenus en amont de la phase de décodage d'un message, sauf celui sur la taille des messages qui peut être estimée à partir du graphe de mots produit lors d'une première phase de RAP.

Sur un corpus composé de 102 000 tours de parole extrait de dialogues du corpus HMIHY, nous avons obtenu les 6 classes suivantes, que l'on peut caractériser *a posteriori* comme suit :

- **C1** : message contenant entre 10 et 15 mots en réponse à un prompt demandant une valeur numérique ;
- **C2** : message contenant entre 5 et 10 mots en réponse à un prompt de confirmation ;
- **C3** : message contenant moins de 5 mots en réponse à un prompt de confirmation ;
- **C4** : message contenant moins de 10 mots en réponse à un autre prompt ;
- **C5** : message contenant entre 10 et 15 mots en réponse à un autre prompt ;
- **C6** : message contenant plus de 15 mots en réponse à un autre prompt ;

Il est intéressant de constater que les classes obtenues reflètent deux dimensions : le contexte du dialogue, identifié par le type de prompt joué immédiatement avant le message ; et le langage utilisé, caractérisé par la longueur des messages. Le fait que le contexte de dialogue sélectionné se réduise au prompt précédent s'explique par la faible taille des dialogues sur le corpus HMIHY, qui est une application de routage d'appel.

On retrouve ici la corrélation entre taille des messages et langage utilisé que l'on a aussi constaté sur le corpus France Télécom FT3000 (publié dans l'étude (Damnati et al., 2007a)) : les utilisateurs réguliers produisent des formulations courtes, proche d'une commande par mots clés, la perplexité est donc très faible sur ces messages. A l'inverse les messages longs sont souvent le signe d'utilisateurs novices, ne sachant pas comment exprimer leurs requêtes, et ayant tendance à rajouter des commentaires souvent mal reconnus par la machine. Par exemple la classe des messages répondant à des demandes de confirmation par des messages assez longs (plus de 5 mots) est un bon indicateur de problèmes dans le dialogue.

Ici le contexte de production des messages est donc identifié par deux dimensions, celle provenant du système à travers les prompts et celle provenant de l'utilisateur par rapport au registre de langue utilisé caractérisant son degré de familiarité avec le système.

Pour chaque classe d'états ainsi définis un modèle de langage spécifique a été appris. Lors du décodage le modèle correspondant au contexte est utilisé en conjonction avec un modèle général. Les résultats obtenus avec les modèles adaptés, sur un corpus de tests composé de 7000 tours de dialogue, sont présentés dans le tableau 6.3.

La baisse de perplexité est sensible entre les modèles génériques et ceux adaptés au contexte : baisse de 25.3 à 18.5 sur tout le corpus, soit une baisse relative de 26.8%. Le taux d'erreurs mots baisse lui aussi, mais dans une proportion bien moindre : baisse absolue d'un peu plus de 1% sur tout le corpus, néanmoins un gain est constaté pour chaque classe. Il est intéressant de constater que le gain en perplexité et taux d'erreurs

	Perplexité		WER %	
C	**sans adapt.**	**avec adapt.**	**sans adapt.**	**avec adapt.**
1	18.6	13.9	11.3	11.1
2	5.0	3.2	14.5	12.5
3	3.2	1.5	4.4	2.5
4	11	7.4	19.2	18
5	11.3	9.5	19.7	18.8
6	38.4	27.4	30.8	29.8

TAB. 6.3: *Résultats en perplexité et taux d'erreurs mots (WER) avec et sans adaptation au contexte pour chaque classe de messages sur le corpus HMIHY*

mots concerne aussi la classe 6. Cette classe contient environ 72% des mots du corpus, il s'agit essentiellement des réponses longues au prompt de départ du dialogue : « *Comment puis-je vous aider ?* ». Dans ce cas le modèle adapté consiste essentiellement à diminuer la part des messages très courts et formatés dans les probabilités du modèle de langage général. Cela indique donc qu'avoir plusieurs registres de langue dans le même modèle est source de confusion et que l'on a effectivement intérêt à considérer ces registres comme des informations cruciales pour caractériser le contexte de production d'un message et avoir des modèles dépendant de ce contexte.

6.2.3 Discussion

Les deux expériences rapportées dans cette section démontrent l'intérêt d'adapter les modèles de RAP et de compréhension au contexte de production des messages. Le point faible de ces méthodes est bien évidemment la définition et la caractérisation de ce contexte d'une part et l'attribution de manière dynamique d'une étiquette de contexte à un message donné d'autre part. L'approche intégrée RAP/compréhension proposée dans cette étude permet d'utiliser directement les informations contextuelles pour réévaluer les hypothèses produites dans une première phase de décodage. Enfin l'association du registre de langue à la notion de contexte nous parait également un point crucial. Ce point est l'une des motivations principales du projet ANR EPAC d'étude de la parole spontanée dans les corpus d'émissions radiodiffusées qui a commencé en 2007 et dans lequel le LIA est engagé.

Enfin notons que dans le cadre du dialogue oral, nous avons également étudié l'intégration de la notion de contexte de dialogue directement dans la phase de décodage, en gérant un espace d'états possibles à chaque tour de dialogue. Cette étude a été menée à la fois sur le corpus MEDIA dans le cadre de la campagne d'évaluation sur la spécification du sens en contexte de dialogue publiée à la conférence TALN'07 (Denis et al., 2007) ; et aussi sur le corpus FT3000 dans un article publié à ICASSP'07 (Damnati et al., 2007a).

6.3 Mesures de confiance, stratégie d'interprétation et de correction

La production d'un espace de recherche sur les interprétations, représenté par le transducteur mot/concept décrit au chapitre 4 et la liste structurée de n-meilleures interprétations, permet au module utilisant les sorties du système de compréhension de définir des stratégies prenant en compte cet espace plutôt qu'une hypothèse unique. Ces stratégies peuvent s'attacher à réévaluer les hypothèses produites en fonction de la tâche à accomplir ou à définir des mesures de confiance conditionnant la suite du processus, telles que des demandes de confirmation ou de clarification dans le cadre d'un dialogue. Nous allons présenter deux exemples de stratégies : l'une visant à obtenir des mesures de confiance spécifiques au processus de compréhension, développée sur le corpus de France Télécom PLANRESTO ; l'autre portant sur une stratégie d'exploitation des différents sorties des processus de RAP et de compréhension (meilleure hypothèse, liste d'hypothèses, graphe d'hypothèses, réseau de confusion) appliquée aux corpus HMIHY et FT3000.

6.3.1 Mesures de confiance par le consensus de classifieurs

Les modèles utilisés pour effectuer le décodage conceptuel présenté au chapitre 4 sont des modèles génératifs : modèles de langage n-grammes et modèles d'étiquetage à base de modèles de Markov cachés. Une fois le décodage effectué, des modèles discriminants de type classifieurs (par exemple les SVM) ou étiqueteurs (par exemple les CRF) peuvent être employés sur les meilleurs hypothèses produites pour réévaluer les interprétations ou spécifier certaines propriétés des concepts comme présenté dans (Denis et al., 2007) sur le corpus MEDIA.

Une autre utilisation possible des méthodes discriminantes est la possibilité de donner une décision (ou une probabilité par régression) sur l'exactitude d'une hypothèse, à partir des valeurs données par les différents modèles au cours du décodage : mesures de confiance acoustique, linguistique et sémantique, probabilité a posteriori, contexte. L'estimation de cette décision ou de la probabilité $P(correct(\Gamma)|\Phi)$, c'est à dire la probabilité que l'interprétation Γ soit correcte étant donné un ensemble de mesures de confiance Φ, peut être apprise sur un corpus de développement contenant à la fois toutes les mesures de confiance pour chaque message mais aussi l'étiquette de référence permettant de déterminer si une hypothèse est correcte ou pas. On se ramène donc ici à un problème de classification binaire quelle que soit la complexité de la représentation sémantique utilisée.

Nous avons publié une telle méthode dans un article de la revue *IEEE Transaction on Speech and Language Processing* (Raymond et al., 2007). Dans cet article nous utilisons le critère de *redondance* de la décision entre plusieurs classifieurs comme mesure de confiance. On définit ainsi une *unité de décision* comme une mesure de l'accord entre plusieurs processus de décision prenant en entrée différentes représentations du

condition	*All*	$DU_1 \wedge DU_2$	$DU_1 \wedge \overline{DU_2}$	$\overline{DU_1} \wedge DU_2$	$\overline{DU_1} \wedge \overline{DU_2}$
couverture	*100%*	58.4%	16.2%	7.6%	17.7%
CER	17.0	5.9	28.6	18.0	30.8
NCE	-	**0.27**	0.09	0.00	0.07

TAB. 6.4: *Taux d'erreurs sur les concepts (CER), couverture et réduction de l'entropie croisée normalisée (NCE) en fonction des états de confiance définis par les unités de décision DU_1 et DU_2*

même phénomène et devant prédire une étiquette commune. Par exemple sur le corpus PLANRESTO deux unités de décision ont été définies : l'unité DU_1 qui vérifie la cohérence de l'interprétation sémantique d'un message et l'unité DU_2 qui prédit sa fiabilité par rapport aux mesures de confiance obtenues durant la phase de décodage (RAP+décodage conceptuel).

Trois classifieurs sont utilisés : un classifieur à base d'arbres de décision, proposé pour l'interprétation sémantique par (Kuhn et De Mori, 1995), dans l'implémentation décrite dans (Béchet et al., 2000) ; et deux classifieurs à large marge, un fondé sur les SVM de Vapnick dans l'implémentation de (Collobert et al., 2002), et un classifieur implémentant un algorithme de *boosting* de classifieurs simples, *BoosTexter* proposé par (Schapire et Singer, 2000).

Voici la rapide description de ces unités de décision :

- Dans l'unité DU_1 la meilleure interprétation produite durant la phase de décodage conceptuel est traitée par trois classifieurs. Les paramètres donnés en entrée aux classifieurs sont les séquences de mots et de concepts reconnus. Chaque classifieur construit ensuite sa propre représentation : expressions régulières pour les arbres de décision, sac de mots pour les SVM et sac de n-grammes pour BoosTexter. L'unité DU_1 est validée en fonction de l'accord des trois classifieurs sur l'interprétation produite.
- L'unité DU_2 utilise également les mêmes classifieurs ainsi que la règle d'accord, mais cette fois les paramètres donnés en entrée sont les scores de confiance obtenus lors de la phase de décodage : confiance acoustique, linguistique, sémantique, rang de l'hypothèse dans la liste d'hypothèse structurée et enfin confiance par rapport au contexte en utilisant une distribution *a priori* des interprétations pour un contexte donné.

En fonction de la validation de ces unités de décision, différents états de confiance sont définis. Ces états de confiance ont deux utilités : d'une part influencer la gestion du dialogue en permettant d'omettre ou au contraire de rajouter une phase de confirmation après la reconnaissance d'un concept ; d'autre part permettre d'exploiter la liste de n-meilleures hypothèses pour effectuer de la correction d'erreurs dans le cas ou l'hypothèse produite reçoit un niveau de confiance faible du point de vue des unités de décision DU_1 et DU_2.

Les résultats obtenus sur le corpus PLANRESTO sont présentés dans le tableau 6.4. Ces résultats montrent que ces unités de décision sont d'intéressantes mesures de confiance : dans l'état le plus fiable ou les deux unités sont validés, le taux d'erreurs sur les concepts

décroît de 17 à 5.9%, alors que dans l'état considéré comme le moins fiable, ce taux atteint les 30.8%. Une mesure intéressante pour évaluer une mesure de confiance est la mesure NCE (pour Normalized Cross Entropy) proposée par NIST pour la campagne d'évaluation HUB-5 (NIST, 2001).

La plus forte valeur de NCE est obtenue lorsque les deux unités de décision sont validées, un gain significatif d'information est donc apporté dans ce cas là, représentant 58.4% des messages.

6.3.2 Stratégie d'interprétation

Expériences sur le corpus FT3000

Dans les chapitres 4 et 5 j'ai présenté une approche intégrée des processus de RAP et de compréhension exploitant un espace de recherche riche, les graphes de mots produits par une première phase de RAP, et produisant ensuite un espace de recherche d'interprétation. Cette méthode est efficace lorsque l'espace de recherche initial contient effectivement la bonne interprétation, mesurable grâce au score Oracle comme présenté dans la section 6.1.2. Cependant, si cet espace est trop riche, et surtout s'il ne contient pas la bonne interprétation, le risque est grand que l'approche intégrée cherche «*à tout prix*» à produire une interprétation, entraînant par là un risque important de mauvaise interprétation.

Cette situation est étudiée, sur le corpus FT3000, dans un article présenté récemment à la conférence Interspeech à Anvers (Minescu et al., 2007). Ce corpus, présenté dans le chapitre 3, contient des traces de dialogue entre un système mis en service et de vrais utilisateurs. Tous les problèmes concrets tels que les communications interrompues, le bruit dans les messages, les utilisateurs non coopératifs voire furieux, auxquels une application déployée est confrontée, se retrouvent dans ce corpus.

Dans cet article les messages du corpus FT3000 sont classés en quatre catégories :

1. $C1$ contient les messages vides, c'est à dire contenant du bruit sans parole ;
2. $C2$ contient les énoncés hors-domaine, c'est à dire ne contenant aucune information relative au service, mais plutôt des commentaires généraux ou encore des appréciations ;
3. $C3$ contient les énoncés dont le thème correspond bien à celui du service, mais dont la requête n'est pas couverte par les règles d'interprétation ;
4. $C4$ contient les énoncés pertinents, dont les requêtes sont bien couvertes par le service.

La table 6.5 indique les proportions des différentes catégories avec des exemples de messages, sur un corpus collecté sur une période de deux semaines par le service FT3000. Ce corpus représente une «*photo réaliste*» des messages que doit traiter une application mise en service.

Cat.	**nb**	**%**	**exemple**
*C*1	1333	20.5%	« *biiiiiip* »
*C*2	674	10.4%	« *bon bon qu'est ce que je dois dire là euh euh* »
*C*3	355	5.5%	« *on m'appelle tout le temps y'a personne je sais pas qui c'est* »
*C*4	4139	63.7%	« *je voudrais payer ma facture* »

TAB. *6.5:* *Différentes catégories de messages sur le corpus FT3000*

cat.	*C*1	*C*2	*C*3	*C*4
erreur	FA	FA	FA	FR+Sub
séquentielle	6.5%	7.8%	2.9%	8.8%
intégrée	22.8%	13.0%	6.3%	6.5%

TAB. *6.6:* *Taux d'erreurs obtenus sur chaque catégorie de message pour les deux approches*

Comme on peut le voir, 36.3% des messages sont dans les classes *C*1, *C*2, et *C*3, c'est à dire que 36.3% des messages devraient être rejetés car non pertinents pour le service. Ce point illustre parfaitement la différence entre les corpus *écologiques* ou extraits d'applications mises en service, et les corpus de laboratoire de type MEDIA ou ATIS. Dans ces derniers, seulement les messages de la catégorie *C*4 sont considérés. Ainsi il est primordial lors de tout transfert d'une application de laboratoire à une application industrielle d'évaluer les modèles proposés également sur des messages *a priori* à rejeter tels que les messages des catégories 1 à 3, dans la mesure où ils représentent une part non négligeable du trafic d'un système déployé.

Nous avons alors comparé deux approches sur les messages des différentes catégories : l'approche séquentielle consistant à n'analyser que la meilleure séquence de mot produite par le module de RAP et l'approche intégrée utilisant un graphe de mots pour chercher conjointement les interprétations et les séquences de mots. L'évaluation est faite avec les mesures suivantes :

- *Fausse Alarme* (FA), quand une interprétation est produite à partir d'un message M à rejeter (avec $M \in \{C1, C2, C3\}$) ;
- *Faux Rejet* (FR), quand un message M est rejeté alors qu'il contenait une interprétation (avec $M \in \{C4\}$) ;
- *Substitution* (Sub), quand l'hypothèse d'interprétation produite pour le message M est différente de celle contenue dans la référence (avec $M \in \{C4\}$).

Comme nous pouvons le voir dans la table 6.6, si l'approche intégrée permet bien d'améliorer les performances de compréhension pour les messages pertinents de la catégorie *C*4, en revanche l'utilisation de graphes de mots en entrée du module de compréhension pour les messages bruités provoque une augmentation très importante du taux de Fausse Alarme.

De manière caricaturale ce taux de FA bondit de 6.5% à 22.8% pour les messages *C*1 ne contenant pas de parole. Cela s'explique par le fait que si aucun mécanisme de rejet utilisant les scores de vraisemblance acoustique n'est appliqué, le module de RAP va projeter n'importe quel signal vers un espace de mots, quel que soit le signal pris

cat.	*C*1 + *C*2 + *C*3	*C*4		*C*1 + *C*2 + *C*3 + *C*4
erreur	FA	FR	Sub	IER
strat1	17.2%	2.7%	6.1%	26.0%
strat2	8.8%	5.2%	4.1%	18.1%

TAB. 6.7: *Taux d'erreurs obtenus avec la stratégie **strat2** n'appliquant l'approche intégrée que sur les messages fiables au sens de plusieurs mesures de confiance. **strat1** correspond à la méthode séquentielle de base traitant la meilleure hypothèse de mots issus des modules de RAP*

en entrée. Sur des signaux contenant pas ou peu de parole cet espace sera très riche car tous les chemins étant uniformément mauvais, l'algorithme de recherche aura du mal à détacher un meilleur chemin de manière claire. Dans un tel espace le module de compréhension pourra toujours trouver une interprétation et ainsi, si aucune mesure de rejet n'est effectuée, produira une fausse alarme expliquant le taux de 22.8% sur *C*1.

Heureusement des mesures de confiance permettent dans la plupart des cas de rejeter directement les messages *C*1 et dans une moindre mesure les messages *C*2. Il n'en va pas de même pour les messages *C*3, contenant de la parole couvrant la même thématique que celle du service déployée, et donc utilisant un langage proche de celui ayant servi à entraîner les modèles linguistiques du module de RAP. Dans le tableau 6.6, ce taux est presque doublé entre l'approche séquentielle et intégrée. Ceci s'explique par la possibilité donnée au module de compréhension de rechercher directement dans le graphe une interprétation, quitte à s'écarter de ce qui a été dit lorsque les scores issus du module de RAP ne sont pas assez discriminants.

Il apparaît donc crucial d'élaborer des stratégies qui n'utilisent pas les même méthodes de compréhension selon le type de message à décoder. Certaines méthodes, comme l'approche intégrée, sont performantes pour les messages *pertinents*, alors qu'elles s'avèrent risquées pour traiter les messages hors-domaines. Détecter aussi tôt que possible ces énoncés hors-domaine est à mon sens une nécessité dans les applications amenées à traiter des messages «*réalistes*». C'est une stratégie de ce type que nous avons présentée dans (Minescu et al., 2007) : la méthode intégrée prenant en compte le graphe complet de mots issus du module de RAP n'est appliquée que sur l'ensemble des messages ayant été considérés comme pertinents selon un ensemble de mesures de confiance. Les autres messages sont soit rejetés, soit traités avec une approche séquentielle pour éviter de cumuler les erreurs au fil de la chaîne de traitement.

Comme nous pouvons le voir dans la table 6.7, la stratégie **strat2** n'appliquant l'approche intégrée que sur les messages *fiables*, permet d'une part de faire diminuer très significativement le taux de fausse alarme, tout en gardant un gain sur le taux de substitution par rapport à l'approche séquentielle (**strat1**). Le fait d'effectuer du rejet dans **strat2** implique une augmentation des erreurs de Faux Rejet, cependant le taux d'erreur d'interprétation globale (Interpretation Error Rate - IER) est grandement amélioré par cette méthode, par rapport à la méthode **strat1** où le rejet est effectué en bout de chaîne par le module de compréhension.

Expériences sur le corpus HMIHY

Un autre exemple de stratégie exploitant le graphe de mots issus de la phase de RAP en fonction du type de message traité est illustré par la figure 6.10 sur le corpus HMIHY pour une tâche de détection et d'extraction d'entités nommées. Cette stratégie, présentée dans un article de la revue *Speech Communication* (Bechet et al., 2004), consiste à détecter d'abord sur la meilleure séquence de mots produite par le module de RAP les **segments** susceptibles de contenir une entité nommée d'un type particulier. Cette détection se fait grâce à un étiqueteur probabiliste fondé sur les Modèles de Markov Cachés. Cet étiqueteur est entraîné sur des transcriptions automatiques du corpus HMIHY et est réglé pour maximiser la mesure de rappel au détriment de la précision.

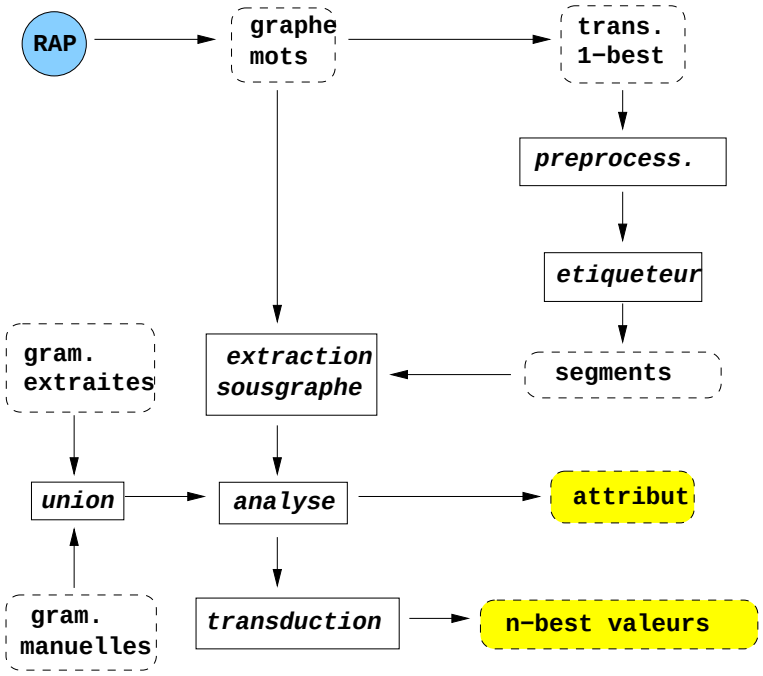


FIG. 6.10: *Stratégie d'extraction des entités nommées pour le corpus How May I Help You ? (HMIHY)*

Pour chaque segment détecté, le sous-graphe correspondant à la même période temporelle, dans le graphe de mots issus de la phase de RAP, est extrait. Les grammaires correspondant au type d'entité nommée détectée sont appliquées à ce sous-graphe, s'il existe au moins une analyse possible le segment est validé et les meilleurs valeurs pour l'entité sont extraites par un transducteur comme présenté dans la section 4.4. Un exemple d'application de cette stratégie pour l'extraction d'entités *numéros de téléphone* est donné en figure 6.11.

La table 6.8 montre les résultats obtenus pour la tâche de détection et d'extraction de valeurs pour les entités *numéros de téléphone* dans le corpus HMIHY. Les corpus d'apprentissage et de test contiennent respectivement 102 000 et 28 000 énoncés. Comme la formation de ces entités obéit à des règles strictes, elles sont naturellement représentées

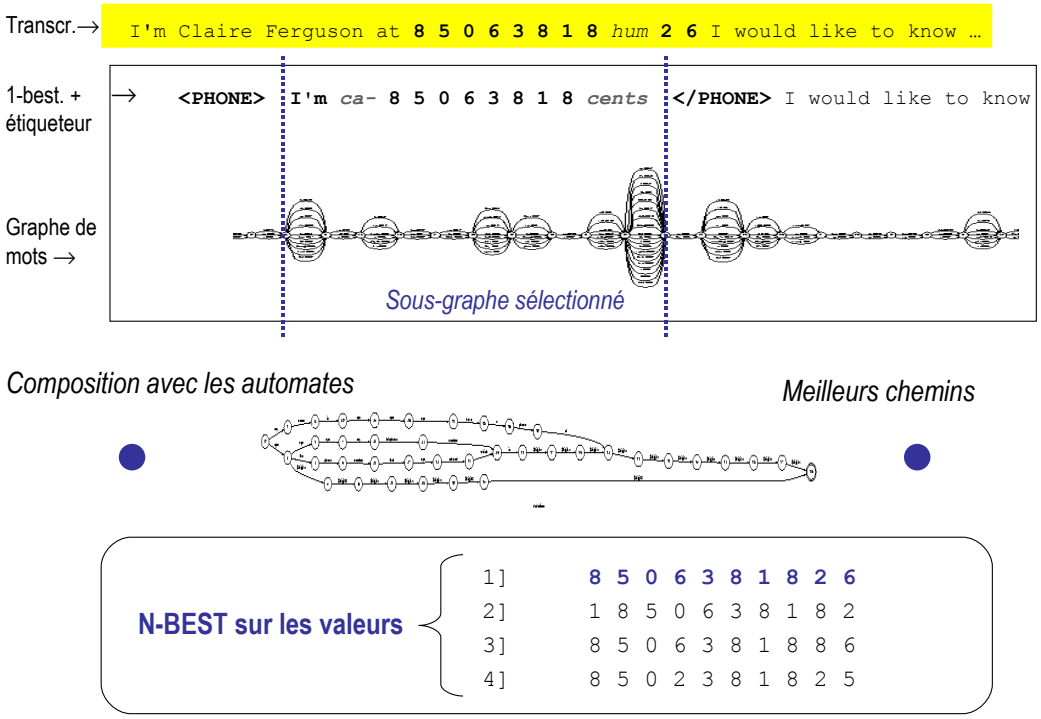


FIG. 6.11: *Exemple de production d'une liste de n-meilleures valeurs lors de la détection d'une entité numéro de téléphone sur le corpus HMIHY*

par des grammaires régulières, comme l'attestent les excellents résultats en précision obtenus par la méthode à base de grammaires. Cependant le corpus HMIHY contient des traces de dialogue avec de vrais utilisateurs, la parole est spontanée et les locuteurs ont souvent tendance à hésiter ou à se reprendre lors de l'énoncé d'un numéro de téléphone à l'intérieur d'un message, comme dans l'exemple de la figure 6.11. Ces disfluences font que la mesure de rappel pour la détection obtenue avec les grammaires sur ces entités est faible : 65.7%. En utilisant conjointement un étiqueteur entraîné spécialement pour maximiser le rappel et les mêmes grammaires mais cette fois appliquées sur les sous-graphes de mots correspondant aux segments détectés, la mesure de rappel augmente considérablement pour atteindre 80.9%.

tâche	**détection entité**		**extraction valeur**	
méthode	**gram. 1-best**	**etiq.+gram.**	**gram. 1-best**	**etiq.+gram.**
Précision	98.5	94.8	88.4	74.1
Rappel	65.7	80.9	59.0	63.3
mesure F	78.9	87.3	70.8	68.3

TAB. 6.8: *Résultats de détection et d'extraction d'entités numéro de téléphone avec des grammaires sur la meilleurs chaîne de mots issue de la RAP et avec la stratégie d'extraction dans les sous-graphes correspondant aux segments détectés par l'étiqueteur en entité nommées*

Ces bons résultats de détection n'entraînent cependant pas de bons résultats pour la tâche d'extraction des valeurs. En effet, si beaucoup plus de numéros sont détectés dans les zones perturbées par des disfluences, la meilleure valeur retrouvée n'est pas souvent entièrement correcte, comme le montre la faible augmentation du rappel, de 59.0% à 63.3%, et la chute brutale de précision de 88.4% à 74.1%. Cependant, en générant des listes de *n*-meilleures valeurs et en utilisant le contexte de l'application, ces performances d'extraction de valeurs peuvent augmenter considérablement. Par exemple, dans (Rahim et al., 2001), il est montré que filtrer une liste de *n*-meilleures hypothèses de numéros de téléphone grâce à un annuaire permet d'obtenir une précision de 94.5% pour les numéros appartenant à cet annuaire contre seulement 45% pour les autres numéros.

Oracle	**1-best**	**2-best**	**5-best**	**10-best**
Précision	74.1	79.5	83.8	84.4
Rappel	63.3	67.9	71.5	72.0
mesure F	68.3	73.2	77.2	77.7

TAB. 6.9: *Performance d'extraction de numéros de téléphone avec la méthode hybride étiqueteur/grammaires en considérant les scores Oracle sur des listes de 1, 2, 5 et 10 hypothèses*

Ainsi la table 6.9 présente les résultats obtenus avec notre méthode en considérant non plus seulement la meilleure hypothèse (1-best), mais des listes de 2, 5 et 10 hypothèses. En gardant seulement 5 valeurs pour chaque numéro détecté la précision et le rappel s'améliorent de près de 10% en absolu.

6.3.3 Discussion

Les résultats présentés dans cette section sur les stratégies de compréhension a mis en lumière deux points clés : d'une part la nécessité d'adapter la méthode de décodage sémantique au message traité ; d'autre part l'intérêt de l'approche intégrée dans un processus de décision prenant en compte l'ensemble des mesures de confiance obtenues lors du décodage ainsi que le contexte de production du message.

Le premier point, l'adaptation des méthodes aux messages, s'explique par la nature des modèles employés, essentiellement des modèles statistiques. Comme indiqué dans le paragraphe 6.2, ces modèles font comme hypothèse fondamentale la cohérence entre les conditions d'apprentissage et de test. Les énoncés hors-domaines brisent cette hypothèse, car ils sont par nature inattendus. Si un processus de détection et de rejet n'est pas mis en place le plus tôt possible dans la phase de traitement, ces fausses alarmes se retrouvent dans la chaîne de décodage. Dans ce cas certaines stratégies sont plus risquées que d'autres. Notamment la recherche d'une interprétation dans un graphe de mots trouvera pratiquement toujours une solution, et il est plus difficile de rejeter cette solution en bout de chaîne que lors des phases initiales du décodage.

Le deuxième point illustre la nécessaire imbrication du processus de compréhension avec l'application utilisant ses sorties (gestionnaire de dialogue ou système de recherche d'information). En définissant de nouvelles mesures de confiance pour chaque

hypothèse produite, le nombre d'interprétations probables conservées dans la liste structurée d'hypothèses d'interprétation peut être réduite. En confrontant cette liste au contexte d'utilisation des messages, le processus de décision peut améliorer notablement les performances de compréhension.

Par exemple sur le message suivant, extrait du corpus HMIHY :// «*I wanna know why I was charged on September sixth 11 dollars 63 cents for calling 8 5 6 2 1 6 5 5 2 1 Clementon New Jersey for 1 minute*»// le locuteur énonce cinq entités nommées : *September sixth*, *11 dollars 63 cents*, *8 5 6 2 1 6 5 5 2 1*, *Clementon New Jersey*, *1 minute*.

Or ces cinq entités sont connues du système de dialogue, dans la mesure où elles apparaissent dans la facture téléphonique du locuteur ; elles sont donc accessibles parmi toutes les informations spécifiques associées au locuteur une fois celui-ci identifié. Ces informations constituent le contexte utilisateur, elles doivent être utilisés pour définir l'ensemble des interprétations attendues.

6.4 Traitement de la parole très spontanée

Plusieurs corpus utilisés dans cette étude contiennent de la parole très spontanée regroupant l'ensemble des phénomènes posant problème aux systèmes de RAP actuels : parole téléphonique, bruits additifs (locuteurs téléphonant dans la rue, leur voiture, etc.), bruits convolutifs (téléphones portables), fort niveau de disfluences (parole non préparée), énoncés hors-domaine (services vocaux ouverts à une très grande échelle, utilisateurs novices ou non coopératifs). Ce catalogue de problèmes rend ces messages particulièrement intéressants à traiter car ils permettent d'envisager les limites des méthodes actuelles de RAP.

6.4.1 Parole spontanée dans le corpus d'opinions de France Télécom

Le corpus d'enquêtes d'opinion de France Télécom est un bon exemple de ce genre de corpus. La table 6.10 présente la transcription d'un message assez verbeux issu de ce corpus, qui sans faire partie des messages les plus difficiles à traiter, illustre bien le type de difficultés auxquelles nous sommes confrontés.

Ce message est un exemple type de parole spontanée ou *non préparée*. D. Luzzati donne une définition précise de ce type de parole : « un énoncé conçu et perçu dans le fil de son énonciation » (Luzzati, 2004). Les nombreuses reprises et corrections contenues dans celui-ci illustrent bien cette définition, et entraînent d'énormes difficultés de transcription de la part des systèmes de RAP. Même en supposant que les modèles de RAP soient suffisamment robustes pour transcrire ces messages en mots, ce type de sorties n'est pas forcément exploitable par les systèmes chargés d'en effectuer la compréhension. En effet dans un énoncé spontané la transcription en mots ne représente qu'une dimension du message. Il manque les informations liées au signal de parole lui-même : prosodie, expressivité, qualité de la voix. Sans ces informations le message

« euh bonjour donc c' est XX à l' appareil je sais pas si vous savez très bien qui je suis euh donc par rapport au à la niveau de la satisfaction de ma satisfaction personnelle par rapport à votre service euh je dirai que dans l' ensemble je suis euh plutôt satisfait euh vous avez un très bon service clientèle qui sait écouter qui euh non qui j' ai pas grand chose à dire c' est c' est très très bien sinon ben juste par rapport au à ce que vous avez mis en place euh tout de suite justement c' est une très bonne idée justement d' une façon à ce qu' y ait un taux de réponse euh assez important maintenant c' est vrai qu' on est obligé de rappeler plusieurs fois et encore quand on prend le temps de rappeler pour euh pour euh pour euh pour répondre parce que quand on nous dit euh vous allez nous donner vous allez donner euh votre euh vos idées euh vos vos suggestions et ben on n' a rien en tête donc c' est pour ça que j' ai été obligé de raccrocher et de réfléchir à ce que je vais vous dire on ça c' est pas je pense euh que ça c' est le point collectif ou c' est le point négatif et sinon dans l' ensemble je suis très satisfait sinon y a une chose que j' ai annoter euh j' ai deux comptes chez vous euh je trouve ça un peu embêtant de pouvoir euh de pas pouvoir accéder euh aux deux par la même personne quand j' appelle mon service clientèle donc ça je trouve ça un peu dommage que je sois obligé de dépenser en plus parce que faut que je c' est pas le même type euh c' est pas la même personne qui s' occupe de mon dossier donc ce qui aurait été bien c' est quand même regrouper les deux dossiers sous euh euh sous un seul quoi de façon à ce que quand on appelle on puisse accéder aux deux dossiers séparément bien sûr mais les deux dossiers donc voilà euh sinon ben je vous remercie en tous cas pour euh pour votre gentillesse et votre amabilité vos conseillers clientèle sont très très gentils et très à l' écoute et donc je vous en remercie au revoir bonne journée bonne soirée »

TAB. *6.10:* *Exemple de message du corpus d'opinions France Télécom*

devient difficilement compréhensible, et il est de toute manière inadapté à toute technique d'analyse profonde, qu'elle soit syntaxique ou sémantique.

En l'absence de représentation fiable d'informations liées au signal, l'analyse par défaut consiste à aller chercher dans l'énoncé les *pépites* d'informations qui permettront de le caractériser et de répondre à la tâche visée. Par exemple, dans le message de la table 6.10, les *pépites* liées à la tâche de détection d'opinions sont :

- *dans l' ensemble je suis euh plutôt satisfait*
- *un très bon service clientèle qui sait écouter*
- *c' est très très bien*
- *c' est le point négatif*
- *dans l' ensemble je suis très satisfait*
- *je trouve ça un peu embêtant*
- *je trouve ça un peu dommage*
- *je vous remercie en tous cas pour euh pour votre gentillesse et votre amabilité*
- *vos conseillers clientèle sont très très gentils et très à l' écoute*
- *je vous en remercie*

C'est à partir de ces éléments qu'une caractérisation du message pourra être faite, par exemple par des méthodes de classification automatique prenant en entrée les segments détectés. Si ces segments peuvent être relativement bien modélisés car se trouvant en nombre dans les messages du corpus d'apprentissage, il n'en va pas de même des autres parties de messages, caractérisés par une très grande variabilité, et mal pris en compte par les modèles statistiques à cause de cette variabilité.

En effet, du fait du degré de liberté laissé aux utilisateurs dans l'énoncé de leur message, on observe une assez grande dispersion dans la distribution des fréquences des mots. Ceci est d'autant plus le cas dans les portions des messages où les utilisateurs relatent l'origine de leur problème qui peut être de nature assez variée. Une fois les noms propres filtrés, le corpus d'apprentissage dans son ensemble contient 2981 mots différents pour un nombre total de 51056 occurrences. Près de la moitié des mots n'apparaissent qu'une seule fois dans le corpus d'apprentissage, et la restriction du lexique aux mots d'occurrence supérieure ou égale à 2, conduit à un lexique de 1564 mots pour un taux de mots hors-vocabulaire égal à 2,8%.

Dans un article présenté à Interspeech'2006 à Pittsburgh (Camelin et al., 2006), un modèle de langage thématique a été introduit pour remédier au problème de la mauvaise modélisation des parties de messages hors-sujet par rapport à la tâche. L'idée est de ne modéliser explicitement que les portions de messages porteuses d'opinion. Pour cela, un sous-corpus a été extrait pour chaque étiquette qui regroupe l'ensemble des segments associés à cette étiquette dans le corpus d'apprentissage initial. Un sous-modèle de langage a ainsi été estimé pour chaque étiquette à partir du sous-corpus associé. Par ailleurs, un modèle englobant de type bigramme portant sur les étiquettes elles-mêmes a été estimé pour modéliser les enchaînements entre les différents segments d'opinion. Les portions qui ne correspondent à aucune expression d'opinion sont quant à elles modélisées par une boucle de phonèmes en contexte, sans contraintes *a priori* sur les enchaînements de phonèmes. Enfin, un sous-modèle supplémentaire a

été estimé pour les segments qui correspondent à des formules de politesse, souvent en début et en fin de message. En effet ces segments présentent une forte régularité et leur modélisation permet d'éviter une trop grande dérive du décodage dans le modèle de phonèmes bouclé. L'ensemble est compilé au sein d'un unique modèle présenté dans la figure 6.12.

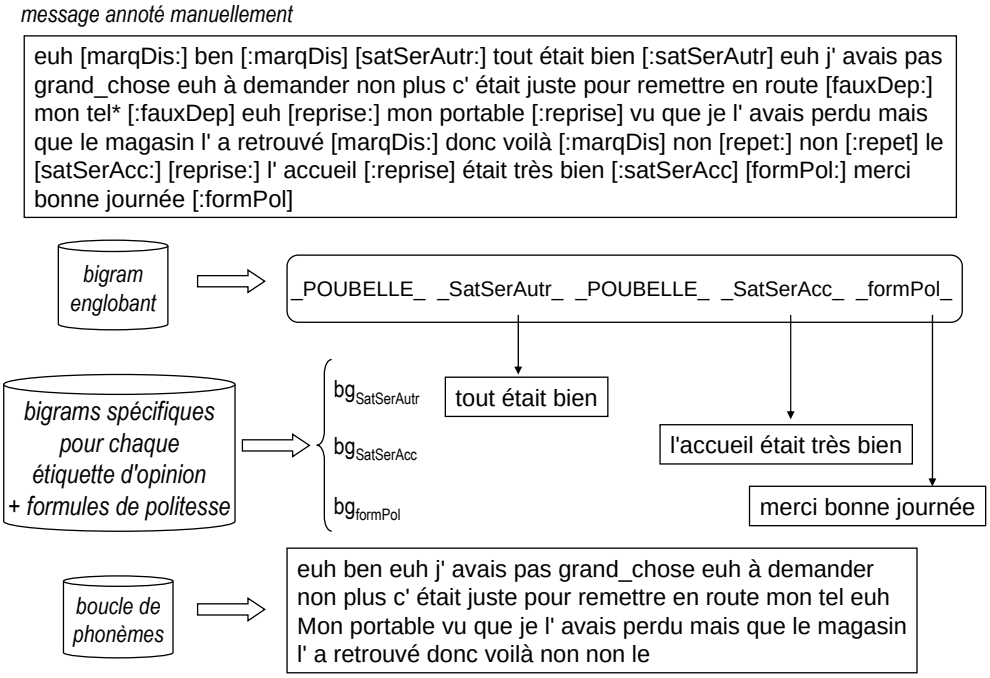


FIG. 6.12: *Exemple de décodage avec les 3 types de modèles de langage thématiques utilisés*

L'ensemble des segments extraits sur toutes les étiquettes représente environ 18700 occurrences de mots et le nombre de mots différents par sous-corpus ne dépasse pas 780 pour une moyenne de 470. Le premier intérêt est donc d'avoir réduit fortement le champ lexical. Par ailleurs, les messages se caractérisent globalement par un haut degré de disfluences. Or à nouveau, les parties les plus disfluentes ne sont pas celles où le locuteur exprime son opinion mais plutôt celles où il relate l'origine de son problème initial. On observe ainsi une réduction du degré de disfluences dans les segments extraits. Ceci est illustré dans le tableau 6.11.

Hormis les répétitions, qui ne sont pas les phénomènes les plus problématiques pour la reconnaissance, l'ensemble des indicateurs ont un pourcentage plus faible dans les segments d'opinion extraits. La baisse la plus significative concerne les marqueurs discursifs qui sont assez difficiles à modéliser du fait de la variété de leurs contextes d'apparition et qui peuvent perturber le traitement ultérieur des messages du fait de leur ambiguïté. Les mots *bon* ou *bien* par exemple peuvent à la fois être porteurs de sens pour une opinion et neutres quand ils sont employés pour articuler le discours.

Indicateur	# messages	# segments
pauses remplies	6.1	5.0
faux départs	1.9	1.7
reprises	4.2	3.9
répétitions	2.0	2.3
marqueurs discursifs	4.3	1.2

TAB. *6.11:* *Pourcentage des indicateurs de disfluences dans le corpus global et dans le corpus extrait*

6.4.2 Parole spontanée pour le corpus FT3000

La parole spontanée ou non préparée, telle que définie dans la section précédente, se retrouve dans le corpus FT3000 dans les commentaires que font les utilisateurs au fur et à mesure de l'interaction, comme présenté dans le chapitre 3. Dans une publication à la conférence ICASSP 2007 (Damnati et al., 2007a), un modèle de langage spécifique à la détection de ces opinions a été proposé, ce modèle s'inspirant des modèles de langage d'opinions mis au point sur le corpus de sondage.

Le point de départ consiste à identifier, dans le corpus d'apprentissage, les segments contenant de tels commentaires des autres segments contenant des énoncés pertinents du point de vue de l'application. Le modèle de langage sur les commentaires contient un vocabulaire de 765 mots, il est entraîné sur un corpus de 1712 segments. Ce modèle, appelé LM^{com}, est mélangé avec un modèle de langage général appelé LM^{G}. Pour apprendre ce modèle de langage général les séquences commentaires ont été supprimées du corpus d'apprentissage et remplacées par le symbole $\$COM$, qui est donc inclus parmi les autres probabilités des n-grammes, suivant le principe des classes de mots *a priori*. Durant le processus de décodage les probabilités du modèle de langage général et du modèle commentaires sont combinées de la manière suivante.
Si P^G est la probabilité du modèle général et P^{COM} la probabilité de LM^{COM}, alors la probabilité de la suite de mots w_1, w_2, w_3, w_4 où la séquence w_2, w_3 est une séquence candidate pour être considérée comme un commentaire est obtenue par :

$$\begin{aligned} P^{G+COM}(w_1,w_2,w_3,w_4) = \quad & P^G(w_1|start) \times P^G(\$COM|w_1) \times \\ & P^{COM}(w_2|start) \times P^{COM}(w_3|w_2) \times \\ & P^{COM}(end|w_3) \times P^G(w_4|\$COM) \end{aligned}$$

Ces modèles ont été entraînés sur un corpus de 44000 messages transcrits et annotés manuellement. Les résultats sont obtenus sur un corpus de tests contenant 1953 messages repartis en 1219 messages étiquetés *transit* et 734 étiquetés *autre*, comme présenté dans le chapitre 3. Les résultats en terme d'erreur d'interprétation (IER), une fois la phase de composition sémantique réalisée, sont présentés dans la table 6.12.

Ces résultats montrent que détecter les commentaires le plus tôt possible, dès la première phase de reconnaissance, permet de filtrer efficacement les messages et d'éviter que les modules ultérieurs ne cumulent des erreurs faites en début de chaîne. Le gain

IER	**tous**	**autre**	**transit**
taille	1953	734	1219
$\mathbf{LM}^{G}$	16.5	22.3	13.0
$\mathbf{LM}^{G+COM}$	15.0	18.6	12.8

TAB. *6.12:* *Taux d'erreur d'interprétation en fonction du modèle de langage utilisé : modèle général seul ou modèle intégrant à la fois un modèle général et un modèle spécifique aux commentaires*

en performance est surtout sensible sur les messages classés *autre*, en effet ce sont eux qui contiennent l'essentiel des commentaires du corpus.

6.4.3 Discussion

L'accès à des corpus *écologiques* contenant de la parole très spontanée est une des particularités des études présentées dans cette habilitation. Ce type de parole met en relief les limites des techniques actuelles et oblige à définir des méthodes pour évaluer la robustesse des hypothèses produites. Même si la réponse apportée est le plus souvent à la fois une restriction des capacités de compréhension à un cadre applicatif très limité et aussi une méthode de rejet permettant d'éliminer les énoncés les plus difficiles, il n'en reste pas moins que ce type de message représente l'un des domaines les plus intéressants des recherche en TAL.

En effet, même avec l'hypothèse forte que l'étape de transcription soit performante (ce qui n'est pas le cas sur ce type de messages), ces transcriptions ne contiennent qu'une partie de l'information permettant d'interpréter un message tel que celui présenté dans la table 6.10. C'est sur ce point que les processus de compréhension de la parole, tels que définis dans ce document, diffèrent de ceux liés à la compréhension du texte écrit, par la nécessaire prise en compte des dimensions acoustiques des messages. Cette prise en compte est un des défis à relever pour la recherche en compréhension de l'oral.

Chapitre 7

Conclusions et Perspectives

Sommaire

Pour conclure cette présentation de mes travaux concernant la problématique de la compréhension automatique de la parole, j'énoncerai quatre principes qui m'apparaissent fondamentaux pour à la fois définir ce domaine de recherche et aussi servir de recommandation à la mise au point de systèmes s'attaquant à cette tâche. Pour chacun de ces principes je rappellerai quels résultats significatifs nous avons obtenus dans les travaux servant de support à cette étude et j'indiquerai quelles sont, à mon avis, les voies de recherche les plus prometteuses les concernant.

7.1 La compréhension de la parole ne se résume pas à une tâche de dictée vocale

Les systèmes de RAP suivent tous l'approche consistant à considérer le traitement de la parole comme une tâche de dictée vocale, la transcription en mots étant le préalable à tout traitement ultérieur. Cette situation s'explique par le fait que la tâche de dictée vocale cadre parfaitement avec le paradigme de la RAP probabiliste : parole préparée, donc proche de la langue écrite sur laquelle sont appris les modèles de langage, d'où validation de l'hypothèse fondamentale d'adéquation entre les données d'apprentissage et d'exploitation. La communication homme-machine ne se résume cependant pas à la seule dictée vocale, et le décalage entre modèles et données est inévitable. Pour

gérer ce décalage les deux voies possibles sont l'adaptation d'une part et la gestion de l'ambiguïté d'autre part. L'adaptation a pour but de rapprocher les modèles des données à traiter, le problème étant la caractérisation de ces nouvelles données et les possibilités limitées d'adaptation des modèles linguistiques. La gestion de l'ambiguïté consiste à garder un espace de recherche ouvert sur les différentes interprétations possibles d'un message. Des stratégies de décision adaptées à la tâche visée peuvent exploiter cet espace, et l'utiliser pour renverser le paradigme de la RAP « traditionnelle » : plutôt que de transcrire d'abord en mots un message pour produire toutes les interprétations possibles de chacune des chaînes de mots produites, les interprétations possibles sont directement recherchées dans l'espace de recherche et à chacune d'elle est associée sa meilleure séquence de mots supports au sens des modèles de RAP.

Le modèle intégré de RAP et de compréhension présenté dans ce document répond à ce principe en produisant cet espace d'interprétation à partir d'un graphe de mots ; les différentes expériences menées montrent que de manière consistante l'approche intégrée obtient de meilleurs scores Oracle que l'approche séquentielle. Cette intégration cependant reste limitée à la production de concepts basiques et de règles d'association sur ces concepts. De plus, une fois le graphe de mots produit, plus aucun retour vers la dimension acoustique des messages n'est effectué hormis la prise en compte de mesures de confiance issues des modèles acoustiques de RAP. Un couplage plus étroit, liant l'interprétation à tous les modèles de RAP, non seulement linguistiques mais aussi acoustiques, est une piste de recherche prometteuse.

7.2 Traiter des transcriptions automatiques ne revient pas à traiter du texte écrit

L'augmentation importante des performances des systèmes de RAP ces dernières années a permis d'envisager cette tâche comme une tâche indépendante, une « boîte noire », qui peut être insérée dans une chaîne de traitement linguistique en permettant de remplacer une entrée textuelle par une entrée vocale. Ainsi on a vu les applications phares du traitement automatique des langues, telles que la traduction automatique, la recherche d'information et le résumé automatique, avoir leur déclinaison vocale : traduction parole-vers-parole (*speech-to-speech*), recherche d'information dans des bases de documents sonores, résumé audio. De plus l'uniformisation des méthodes statistiques, pour l'oral comme pour l'écrit, même pour des tâches *a priori* très linguistiques comme la traduction automatique, font que le format d'entrée des données importe peu pour autant que l'on dispose de corpus d'observations de ces mêmes données en taille suffisante pour apprendre des modèles. Ainsi la tâche *distillation* du récent programme américain DARPA GALE[1] est censée détecter des documents illustrant un certain nombre de scénarios prédéfinis dans des masses de documents qui peuvent être indifféremment sonores ou écrit, voire même des traductions automatiques de documents sonores ou écrit.

[1] http ://www.darpa.mil/ipto/programs/gale/index.htm

Cependant le traitement de transcriptions automatiques plutôt que de textes écrits nécessite la prise en compte de deux différences fondamentales :

- la première concerne la notion d'incertitude ; toutes les hypothèses produites par un système de RAP sont incertaines, cette incertitude pouvant être mesurée par des mesures de confiance ;
- la deuxième est relative au fait que l'espace linguistique des transcriptions automatiques est clos car le lexique et le modèle de langage des modules de RAP sont statiques et limitent cet espace en le circonvenant à celui des données d'apprentissage des modèles.

Le premier point nécessite une adaptation des modèles traitant les transcriptions automatiques afin de prendre en compte, si possible, cette incertitude. Les systèmes destinés à traiter du texte gèrent habituellement les ambiguïtés d'analyse, mais jamais celles liées à la possible remise en question de la fiabilité d'une observation. Le deuxième point illustre une différence fondamentale dans le cahier des charges de systèmes destinés soit à traiter du texte, soit des transcriptions automatiques. Pour les premiers, la capacité des modèles à appréhender de nouvelles observations est cruciale afin de pouvoir prendre en compte à la fois les mots inconnus des modèles, mais aussi les constructions linguistiques inédites. C'est sur cette capacité à *généraliser* des modèles sur de nouvelles observations que s'évaluent les performances des systèmes. Par exemple, pour la tâche de détection d'entités nommées, la principale difficulté consiste à étiqueter les nouveaux noms propres rencontrés afin de déterminer leur catégorie sémantique (personne, lieu, produit, etc.). Pour les seconds, comme mentionné auparavant, l'espace linguistique est clos, le module de RAP ne pouvant prédire d'autres mots que ceux présents dans son lexique de reconnaissance[2]. De fait toute la capacité de généralisation des modèles développés pour traiter du texte devient inutile. Au contraire, les systèmes de traitement des transcriptions automatiques peuvent tirer partie de ce monde clos pour collecter le maximum d'informations *a priori* sur les mots susceptibles d'être reconnus, comme par exemple, toujours pour les entités nommées, toutes les catégories sémantiques avec leurs fréquences pouvant être associées aux différents noms propres des lexiques de reconnaissance.

Le principe précédent plaidait pour une intégration des processus de RAP et de compréhension afin de rendre plus robuste les transcriptions automatiques. Ce principe plaide aussi pour cette même intégration, mais du coté des traitements linguistiques situés après la RAP cette fois, afin d'adapter ceux-ci aux particularités du traitement de la parole. C'est une approche allant dans ce sens que nous avons présenté pour la tâche d'extraction d'entités nommées sur le corpus ESTER.

[2] certaines études ont pour but d'essayer de détecter et dans certains cas réparer (Bisani et Ney, 2005) les erreurs dues à des mots inconnus, mais aucune solution générique n'existe pour le moment pour résoudre ce problème

7.3 Un message vocal n'est interprétable qu'en fonction de son contexte de production

Comme évoqué précédemment l'inévitable décalage entre modèles et données peut être compensé par une adaptation de ces modèles au *contexte de production* des messages vocaux. Cette notion de contexte est très large, elle englobe à la fois le cadre applicatif et le contexte lié au locuteur lui même. Le cadre applicatif peut être global, il définit les thèmes pouvant être abordés et conditionne le choix du lexique de RAP et des concepts de base du module de compréhension. Ce cadre peut aussi être local, particulièrement dans le domaine du dialogue oral, où le contexte du dialogue modifie la liste des interprétations attendues. Les travaux présentés traitant de la détection d'entités nommées à partir d'un corpus d'émissions radiodiffusées et de l'adaptation de modèles au contexte du dialogue sur le corpus HMIHY vont dans ce sens. Les résultats montrent que plus la définition de ce contexte, représenté par des listes d'interprétations attendues, est précise, meilleurs sont les résultats de RAP et de compréhension.

Le contexte lié au locuteur intègre les caractéristiques intrinsèques au locuteur ayant une influence sur la voix, telles que le sexe, l'âge, l'accent et l'état émotionnel, ainsi que le registre de langue utilisé (parole préparée ou non, niveau de langue). Ces dimensions sont prises en compte dans les systèmes de RAP au niveau des modèles acoustiques, où des modèles proches de la voix du locuteur à traiter peuvent être utilisés, notamment dans une deuxième passe de décodage, la première passe ayant servi à caractériser la voix. Ce n'est cependant pas le cas au niveau linguistique, or plusieurs résultats de cette étude ont montré l'intéret de caractériser le plus finement possible le registre de langue. Cette caractérisation permet d'appliquer le plus tôt possible des modèles adaptés au registre détecté, par exemple dans le cas des modèles de langage dépendant du contexte sur HMIHY et des modèles de détection des commentaires sur FT3000.

7.4 Un message vocal ne se limite pas à son seul contenu lexical

Enfin ce dernier point est sans doute celui sur lequel les perspectives de recherche sont les plus vastes. La séquence de mots énoncée dans un message vocal ne constitue qu'une dimension de la compréhension de celui-ci. Les autres dimensions, d'ordre acoustiques, concernent tout d'abord la prosodie, mais aussi la qualité de la voix, sa caractérisation.

La prosodie intervient à plusieurs niveaux : pour désambiguïser le sens d'un message, par exemple en caractérisant le mode (affirmatif, interrogatif, négatif) attaché à une proposition ; mais également pour représenter l'*intention*, le but dans lequel le message est réalisé. Dans cette dimension on trouve bien sûr la notion d'état émotionnel, mais aussi des figures rhétoriques telles que l'ironie.

La caractérisation de la voix, sa qualité, peut nous renseigner sur le sens d'un message en enrichissant la notion de contexte de production présentée dans le point précédent. Par exemple, selon que l'on identifie une voix d'un jeune enfant, d'un adulte ou

d'une personne âgée, le sens d'un message peut évoluer. Cela est particulièrement vrai dans cette étude où le *sens* est défini dans la perspective pragmatique de l'action qu'il va générer de la part de la machine.

La prise en compte de ces dimensions acoustiques est quasiment absente des modèles de RAP et de compréhension de l'oral. Même une dimension aussi incontournable que la prosodie est peu, voire pas du tout utilisée dans la modélisation de la parole. Cela ne veut évidemment pas dire qu'aucune étude n'a essayé de tirer parti d'informations prosodiques pour améliorer le traitement de l'oral, par exemple E. Schriberg (Shriberg et al., 2000; Shriberg et Stolcke, 2004) a étudié l'apport d'une modélisation prosodique pour les tâches de segmentation de la parole continue ou de détection d'actes de dialogues. Cependant les difficultés d'extraction, le manque de robustesse des modèles ainsi que la modélisation implicite de certains paramètres prosodiques dans les modèles de RAP, ont fait que cet apport n'a jamais été clairement établi, en tout cas pas pour la tâche de transcription en mots.

Pourtant je pense que c'est justement le domaine des dimensions acoustiques du problème de la compréhension de la parole qui peut d'une part aider à définir ce champ du traitement automatique des langues en le distinguant de la simple application de modèles textuels à des transcriptions automatiques ; et d'autre part initier de nouvelles recherches sur l'intégration de la prosodie dans les modèles de RAP en justifiant son emploi par l'aspect multidimensionnell de la parole, ne se résumant pas à la seule transcription d'un flux sonore en flux de mots évalué par le critère unique du taux d'erreurs mots.

Liste des illustrations

Liste des tableaux

Bibliographie

(Abney, 1998) S. Abney, 1998. Parsing by Chunks. Dans les actes de *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics COLING/ACL*, Volume 9, 41–47. ACL.

(Alexandersson et al., 1998) J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, et S. M., 1998. *Dialogue Acts in Verbmobil-2 Second Edition*. DFKI, Saarbruucken, Universitat Stuttgart, Technische Universitat Berlin.

(Allauzen et al., 2003) C. Allauzen, M. Mohri, et B. Roark, 2003. Generalized algorithms for constructing statistical language models. Dans les actes de *41st Annual Meeting of the Association for Computational Linguistics (ACL'03), Sapporo, Japan*.

(Allwein et al., 2001) E. Allwein, R. Schapire, et Y. Singer, 2001. Reducing multiclass to binary : a unifying approach for margin classifiers. *The Journal of Machine Learning Research 1*, 113–141.

(Antoine et al., 2003) J.-Y. Antoine, J. Goulian, et J. Villaneau, 2003. Quand le TAL robuste s'attaque au langage parlé : analyse incrementale pour la compréhension de la parole spontanée. Dans les actes de *conference annuelle de l'ATALA - Traitement Automatique des Langues Naturelles - TALN*, Batz-sur-Mer, France, 25–34.

(Baker et al., 1998) C. F. Baker, C. J. Fillmore, et J. B. Lowe, 1998. The berkeley framenet project. Dans les actes de *Proceedings of the 17th international conference on Computational linguistics*, Morristown, NJ, USA, 86–90. Association for Computational Linguistics.

(Barras et al., 2001) C. Barras, E. Geoffrois, Z. Wu, et M. Liberman, 2001. Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication 33*(1-2), 5–22.

(Bechet et al., 2004) F. Bechet, A. Gorin, J. Wright, et D. Hakkani-Tur, 2004. Detecting and extracting named entities from spontaneous speech in a mixed initiative spoken dialogue context : How May I Help Y. *Speech Communication 42*, 207–225.

(Béchet et al., 2002) F. Béchet, A. L. Gorin, J. H. Wright, et D. Hakkani-Tür, 2002. Named entity extraction from spontaneous speech in How May I Help You ? Dans les actes de *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*.

(Béchet et al., 2000) F. Béchet, A. Nasr, et F. Genet, 2000. Tagging unknown proper names using decision trees. Dans les actes de *38th Annual Meeting of the Association for Computational Linguistics, Hong-Kong, China*, 77–84.

(Bechet et al., 2004) F. Bechet, G. Riccardi, et D. Hakkani-Tur, 2004. Mining Spoken Dialogue Corpora for System Evaluation and Modeling. Dans les actes de *Proc. Conference Empirical Methods in Natural Language Processing EMNLP2004*, Barcelona, Spain. Association for Computational Linguistics.

(Bellegarda et Silverman, 2000) J. Bellegarda et K. Silverman, 2000. Toward Unconstrained Command and Control : Data-Driven Semantic Inference. Dans les actes de *Sixth International Conference on Spoken Language Processing*. ISCA.

(Bellegarda, 2004) J. R. Bellegarda, 2004. Statistical language model adaptation : review and perspectives. *Speech Communication 42 Issue 1*, 93–108.

(Benzitoun et Veronis, 2005) C. Benzitoun et J. Veronis, 2005. Problemes d'annotation d'un corpus oral dans le cadre de la campagne easy. Dans les actes de *conference annuelle de l'ATALA - Traitement Automatique des Langues Naturelles - TALN*, Volume 2.

(Bisani et Ney, 2005) M. Bisani et H. Ney, 2005. Open Vocabulary Speech Recognition with Flat Hybrid Models. Dans les actes de *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*.

(Bod, 2000) R. Bod, 2000. Combining semantic and syntactic structure for language modeling. Dans les actes de *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*.

(Bonneau-Maynard et Lefevre, 2005) H. Bonneau-Maynard et F. Lefevre, 2005. A 2+1-level stochastic understanding model. Dans les actes de *Automatic Speech Recognition and Understanding workshop (ASRU)*, Porto Rico.

(Bonneau-Maynard et al., 2005) H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, et D. Mostefa, 2005. Semantic annotation of the french media dialog corpus. Dans les actes de *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa, Portugal.

(Boves et al., 2000) L. Boves, D. Jouvet, J. Sienel, R. de Mori, F. Béchet, L. Fissore, et P. Laface, 2000. ASR for automatic directory assistance : the SMADA project. Dans les actes de *ISCA workshop : ASR2000*.

(Brachman, 1979) R. Brachman, 1979. On the epistemological status of semantic networks. *Associative Networks : Representation and Use of Knowledge by Computers*, 3–50.

(Brachman et Schmolze, 1985) R. Brachman et J. Schmolze, 1985. An overview of the KL-ONE knowledge representation system. *Cognitive Science 9(2)*, 171–216.

(Brown et al., 1990) P. Brown, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, et P. Roossin, 1990. A statistical approach to machine translation. *Computational Linguistics 16*(2), 79–85.

(Bunt, 1995) H. Bunt, 1995. Dynamic Interpretation and Dialogue Theory. *The Structure of Multimodal Dialogue 2*, 139–166.

(Camelin et al., 2006) N. Camelin, G. Damnati, F. Bechet, et R. D. Mori, 2006. Opinion mining in a telephone survey corpus. Dans les actes de *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, Pittsburgh, PA, USA, 1041–1044.

(Carletta, 1996) J. Carletta, 1996. Assessing agreement on classification tasks : the kappa statistic. *Computational Linguistics 22*(2), 249–254.

(Carletta et al., 1997) J. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J. Kowtko, et A. Anderson, 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics 23*(1), 13–31.

(Carrasco et Oncina, 1999) R. C. Carrasco et J. Oncina, 1999. Learning deterministic regular grammars from stochastic samples in polynomial time. *RAIRO (Theoretical Informatics and Applications) 33*(1), 1–20.

(Chappelier et al., 1999) J. Chappelier, M. Rajman, R. Aragues, et A. Rozenknop, 1999. Lattice parsing for speech recognition. Dans les actes de *conference annuelle de l'ATALA - Traitement Automatique des Langues Naturelles - TALN*.

(Charniak, 1997) E. Charniak, 1997. Statistical techniques for natural language parsing. *AI Magazine 18*(4), 33–44.

(Charniak et al., 1993) E. Charniak, C. Hendrickson, N. Jacobson, et M. Perkowitz, 1993. Equations for part-of-speech tagging. Dans les actes de *11th National Conference on Artificial Intelligence*, 784–789.

(Chelba et Jelinek, 2000) C. Chelba et F. Jelinek, 2000. Structured language modeling. *Computer Speech and Language 14(4)*, 283–332.

(Chen et al., 2004) L. Chen, J.-L. Gauvain, L. Lamel, et G. Adda, 2004. Dynamic language modeling for broadcast news. Dans les actes de *In International Conference on Speech and Language Processing*, 1281–1284.

(Chinchor et Robinson, 1998) N. Chinchor et P. Robinson, 1998. Muc-7 named entity task definition. Dans les actes de *Proceedings of the Seventh Message Understanding Conference*.

(Choi et al., 2005) Y. Choi, C. Cardie, E. Riloff, et S. Patwardhann, 2005. Identifying sources of opinions with conditional random fields and extraction patterns. Dans les actes de *Conference on Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, 355–362.

(Chomsky, 1957) N. Chomsky, 1957. Syntactic structures (Trad : Structures syntaxiques, Points Seuil).

(Collobert et al., 2002) R. Collobert, S. Bengio, et J. Mariethoz, 2002. Torch : a modular machine learning software library. Dans les actes de *Technical Report IDIAP-RR02-46, IDIAP*.

(Corazza et al., 1994) A. Corazza, R. De Mori, R. Gretter, et G. Satta, 1994. Optimal probabilistic evaluation functions for search controlled by stochastic context-free grammars. *IEEE Transactions on Pattern Analysis and Machine Intelligence 16*(10), 1018–1027.

(Core et Allen, 1997) M. Core et J. Allen, 1997. Coding dialogs with the DAMSL annotation scheme. *AAAI Fall Symposium on Communicative Action in Humans and Machines*, 28–35.

(Damnati et al., 2007a) G. Damnati, F. Béchet, et R. de Mori, 2007a. Spoken language understanding strategies on the France Telecom 3000 Voice Agency corpus. Dans les actes de *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, HI.

(Damnati et al., 2007b) G. Damnati, F. Bechet, et R. D. Mori, 2007b. Experiments on the France Telecom 3000 Voice Agency corpus : academic research on an industrial spoken dialog system. Dans les actes de *Workshop on Bridging the Gap : Academic and Industrial Research in Dialog Technologies, HLT/NAACL*, 48–55. Association for Computational Linguistics.

(den Os et al., 1999) E. den Os, L. Boves, L. Lamel, et P. Baggia, 1999. Overview of the ARISE project. Dans les actes de *6th European Conference on Speech Communication and Technology*, Volume 4, 1527–1530.

(Denis et al., 2007) A. Denis, F. Béchet, et M. Quignard, 2007. Résolution de la référence dans des dialogues homme-machine : évaluation sur corpus de deux approches symbolique et probabiliste. Dans les actes de *conference annuelle de l'ATALA - Traitement Automatique des Langues Naturelles - TALN*, Volume 1.

(Denis et al., 2006) A. Denis, G. Pitel, et M. Quignard, 2006. A deep-parsing approach to natural language understanding in dialogue system : Results from a corpus-based evaluation. Dans les actes de *LREC 2006*, Genoa, Italy.

(Dhillon et al., 2003) R. Dhillon, S. Bhagat, H. Carvey, et E. Shriberg, 2003. Meeting Recorder Project : Dialog Act Labeling Guide. *Report, ICSI Speech Group, Berkeley, USA*.

(Dupont et al., 2005) P. Dupont, F. Denis, et Y. Esposito, 2005. Links between probabilistic automata and hidden Markov models : probability distributions, learning models and induction algorithms. *Pattern recognition 38*(9), 1349–1371.

(Earley, 1970) J. Earley, 1970. An efficient context-free parsing algorithm. *Communications of the ACM 13*(2), 94–102.

(Eskenazi et al., 1999) M. Eskenazi, A. Rudnicky, K. Gregory, P. Constantinides, R. Brennan, C. Bennett, et J. Allen, 1999. Data Collection and Processing in the Carnegie Mellon Communicator. Dans les actes de *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Volume 6, 2695–2698.

(Fabbrizio et al., 2005) G. Fabbrizio, G. Tur, et D. Hakkani-Tür, 2005. Automated Wizard-of-Oz for Spoken Dialogue Systems. Dans les actes de *Ninth European Conference on Speech Communication and Technology*. ISCA.

(Favre et al., 2005) B. Favre, F. Béchet, et P. Nocéra, 2005. Robust named entity extraction from large spoken archives. Dans les actes de *Conference on Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*, 491–498. Association for Computational Linguistics Morristown, NJ, USA.

(Federico et Bertoldi, 2001) M. Federico et N. Bertoldi, 2001. Broadcast news LM adaptation using contemporary texts. Dans les actes de *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark, 239–242.

(Fillmore, 1985) C. Fillmore, 1985. Frames and the semantics of understanding. *Quaderni di Semantica 6*(2), 222–254.

(Fillmore, 1968) C. J. Fillmore, 1968. The case for case. *Emmon W. Bach and Robert T. Harms, editors, Universals in Linguistic Theory*, 1–88.

(Fiscus et al., 2000) J. Fiscus, W. Fisher, A. Martin, M. Przybocki, et D. Pallet, 2000. NIST evaluation of conversational speech recognition over the telephone : English and Mandarin performance results. Dans les actes de *Proceedings of DARPA Broadcast News Workshop*.

(Gallliano et al., 2005) S. Gallliano, É. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, et G. Gravier, 2005. The ESTER phase ii evaluation campaign for the rich transcription of french broadcast news. Dans les actes de *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal.

(Gildea et Jurafsky, 2002) D. Gildea et D. Jurafsky, 2002. Automatic labeling of semantic roles. *Computational Linguistics 28*(3), 245–288.

(Glass et al., 2000) J. Glass, J. Polifroni, S. Seneff, et V. Zue, 2000. Data collection and performance evaluation of spoken dialogue systems : The MIT experience. Dans les actes de *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*.

(Gorin et al., 1997) A. L. Gorin, G. Riccardi, et J. Wright, 1997. How May I Help You ? Dans les actes de *Speech Communication*, Volume 23, 113–127.

(Gould et al., 1983) J. D. Gould, J. Conti, et T. Hovanyecz, 1983. Composing letters with a simulated listening typewriter. *Commun. ACM 26*(4), 295–308.

(Haffner et al., 2003) P. Haffner, G. Tur, et J. Wright, 2003. Optimizing SVMs for complex call classification. Dans les actes de *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong-Kong.

(Hain et al., 2005) T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, et S. Renals, 2005. Transcription of conference room meetings : an investigation. Dans les actes de *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*.

(Hakkani-Tür et al., 2006) D. Hakkani-Tür, F. Béchet, G. Riccardi, et G. Tur, 2006. Beyond ASR 1-best : Using word confusion networks in spoken language understanding. *Computer Speech & Language 20*(4), 495–514.

(Hakkani-Tur et Tur, 2007) D. Hakkani-Tur et G. Tur, 2007. Statistical Sentence Extraction for Information Distillation. Dans les actes de *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 4.

(He et Young, 2003) Y. He et S. Young, 2003. Hidden vector state model for hierarchical semantic parsing. Dans les actes de *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, 268–271.

(Jackendoff, 1990) R. Jackendoff, 1990. Semantic structures. *The MIT Press, Cambridge Mass.*.

(Jamoussi et al., 2003) S. Jamoussi, K. Smaïli, et J.-P. Haton, 2003. Vers la compréhension automatique de la parole : extraction des concepts par réseaux bayésiens. Dans les actes de *conference annuelle de l'ATALA - Traitement Automatique des Langues Naturelles - TALN*.

(Jermann, 1996) P. Jermann, 1996. Conception et analyse d'une interface semi-structurée dédiée à la co-résolution de problème. *Mémoire de Diplôme d'Etudes Supérieures en Sciences et Technologies de l'Apprentissage. TECFA, Faculté de Psychologie et des Sciences de l'Education, Université de Genève*.

(Jurafsky et al., 1997) D. Jurafsky, E. Shriberg, et D. Biasca, 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. *Institute of Cognitive Science Technical Report*, 97–02.

(Kasami, 1965) T. Kasami, 1965. An efficient recognition and syntax analysis algorithm for context-free languages. *Science Report AFCRL-65-758. Bedford, MA : Air Force Cambridge Research Laboratory*.

(Kelley, 1984) J. F. Kelley, 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst. 2*(1), 26–41.

(Kieffer et al., 2000) B. Kieffer, H.-U. Krieger, et M.-J. Nederhof, 2000. Efficient and robust parsing of word graphs. *W. Wahlster, editor, Verbmobil : Foundations of Speech-to-Speech Translation*, 280–295.

(Kingsbury et Palmer, 2003) P. Kingsbury et M. Palmer, 2003. PropBank : the Next Level of TreeBank. *Proceedings of Treebanks and Lexical Theories*.

(Kuhn et De Mori, 1995) R. Kuhn et R. De Mori, 1995. The application of semantic classification trees to natural language understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence 17*(449-460).

(Lafferty et al., 2001) J. Lafferty, A. McCallum, et F. Pereira, 2001. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Dans les actes de *Proc. 18th International Conf. on Machine Learning*, 282–289. Morgan Kaufmann, San Francisco, CA.

(Lefevre, 2006) F. Lefevre, 2006. A DBN-based multi-level stochastic spoken language understanding system. Dans les actes de *IEEE/ACL Workshop on Spoken Language Technology*.

(Levin et Pieraccini, 1995) E. Levin et R. Pieraccini, 1995. Concept-based spontaneous speech understanding system. Dans les actes de *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Madrid, Spain, 555–558.

(Ljolje et al., 1999) A. Ljolje, F. Pereira, et M. Riley, 1999. Efficient General Lattice Generation and Rescoring. Dans les actes de *Sixth European Conference on Speech Communication and Technology*. ISCA.

(Luzzati, 2004) D. Luzzati, 2004. Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané. Dans les actes de *MIDL*, Paris, France.

(Mangu et al., 2000) L. Mangu, E. Brill, et A. Stolcke, 2000. Finding consensus in speech recognition : Word error minimization and other applications of confusion networks. *Computer, Speech and Language 14(4)*, 373–400.

(Marcus et al., 1994) M. Marcus, B. Santorini, et M. Marcinkiewicz, 1994. Annotated Corpus of English : The Penn Treebank. *Computational Linguistics 19*(2), 313–330.

(Merialdo, 1994) B. Merialdo, 1994. Tagging English text with a probabilistic model. *Computational Linguistics 20*(2), 155–171.

(Miller et al., 2000) D. Miller, S. Boisen, R. Schwartz, R. Stone, et R. Weischedel, 2000. Named entity extraction from noisy input : Speech and OCR. Dans les actes de *Proceedings of ANLP-NAACL 2000*, 316–324.

(Minescu et al., 2007) B. Minescu, G. Damnati, F. Béchet, et R. D. Mori, 2007. Conditional use of word lattices, confusion networks and 1-best string hypotheses in a sequential interpretation strategy. Dans les actes de *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, Antwerp, Belgium.

(Mohri et al., 1997) M. Mohri, F. Pereira, et M. Riley, 1997. AT&T FSM Library - Finite State Machine Library. *AT&T Labs - Research*.

(Mohri et al., 2000) M. Mohri, F. Pereira, et M. Riley, 2000. The design principles of a weighted finite-state transducer library. *Theoretical Computer Science 231*, 17–32.

(Mohri et al., 2002) M. Mohri, F. Pereira, et M. Riley, 2002. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language 16*(1), 69–88.

(Nasr et al., 1999) A. Nasr, Y. Estève, F. Bechet, T. Spriet, et R. D. Mori, 1999. A language model combining n-grams and stochastic finite state automata. Dans les actes

de *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Budapest.

(NIST, 2001) NIST, 2001. The 2001 NIST Hub-5 Evaluation Plan. *http ://www.nist.gov/speech/tests/ctr/h5_2001/h5-01v1.1.pdf*.

(Olive, 2005) J. Olive, 2005. Global autonomous language exploitation (GALE). DARPA Information Processing Technology Office program announcement.

(Pallet, 1997) D. Pallet, 1997. Overview of the 1997 DARPA Speech Recognition Workshop. Dans les actes de *Proc. of DARPA Speech Recognition Workshop, Feb*, 2–5.

(Pallett et al., 1992) D. Pallett, N. Dahlgren, J. Fiscus, W. Fisher, J. Garofolo, et B. Tjaden, 1992. DARPA February 1992 ATIS benchmark test results. Dans les actes de *Workshop on Speech and Natural Language*, 15–27. Association for Computational Linguistics Morristown, NJ, USA.

(Paroubek et al., 2005) P. Paroubek, L. Pouillot, I. Robba, et A. Vilnat, 2005. EASy : campagne d'évaluation des analyseurs syntaxiques. Dans les actes de *conference annuelle de l'ATALA - Traitement Automatique des Langues Naturelles - TALN*.

(Peckham, 1993) J. Peckham, 1993. A New Generation of Spoken Dialogue Systems : Results and Lessons from the Sundial Project. Dans les actes de *Third European Conference on Speech Communication and Technology*. ISCA.

(Popescu et Etzioni, 2005) A.-M. Popescu et O. Etzioni, 2005. Extracting product features and opinions from reviews. Dans les actes de *Conference on Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, 339–346.

(Pradhan et al., 2004) S. Pradhan, W. Ward, K. Hacioglu, J. Martin, et D. Jurafsky, 2004. Shallow semantic parsing using support vector machines. Dans les actes de *conference Human Language Technology HLT/NAACL*.

(Rahim et al., 2001) M. Rahim, G. Riccardi, L. Saul, J. H. Wright, B. Buntschuh, et A. L. Gorin, 2001. Robust numeric recognition in spoken language dialogue. Dans les actes de *Speech Communication*, Volume 34, 195–212.

(Ramabhadran et al., 2003) B. Ramabhadran, J. Huang, et M. Picheny, 2003. Towards automatic transcription of large spoken archives - english ASR for the MALACH project. Dans les actes de *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 216–219.

(Ramshaw et Marcus, 1995) L. Ramshaw et M. Marcus, 1995. Text chunking using transformation-based learning. Dans les actes de *Third ACL Workshop on Very Large Corpora*, 82–94. Cambridge MA, USA.

(Ratnaparkhi et al., 1996) A. Ratnaparkhi et al., 1996. A maximum entropy model for part-of-speech tagging. Dans les actes de *Conference on Empirical Methods in Natural Language Processing EMNLP*, 133–142. Association for Computational Linguistics.

(Raymond et al., 2007) C. Raymond, F. Bechet, N. Camelin, R. De Mori, et G. Damnati, 2007. Sequential Decision Strategies for Machine Interpretation of Speech. *IEEE Transactions on Audio, Speech and Language Processing 15*(1), 162–171.

(Raymond et al., 2006) C. Raymond, F. Bechet, R. D. Mori, et G. Damnati, 2006. On the use of finite state transducers for semantic interpretation. *Speech Communication 48,3-4*, 288–304.

(Raymond et Riccardi, 2007) C. Raymond et G. Riccardi, 2007. Generative and discriminative algorithms for spoken language understanding. Dans les actes de *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, Volume 2.

(Raymond et al., 2007) C. Raymond, G. Riccardi, K. J. Rodríguez, et J. Wisniewska, 2007. The LUNA corpus : an annotation scheme for a multi-domain multi-lingual dialogue corpus. Dans les actes de *The 11th Workshop on the Semantic and Pragmatics of Dialogue DECALOG'07*.

(Riccardi et Gorin, 1998) G. Riccardi et A. L. Gorin, 1998. Language models for speech recognition and understanding. Dans les actes de *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, Sidney, Australia.

(Richardson et Domingos, 2006) M. Richardson et P. Domingos, 2006. Markov logic networks. *Machine Learning 62*(1), 107–136.

(Riloff et Wiebe, 2003) E. Riloff et J. Wiebe, 2003. Learning extraction patterns for subjective expressions. Dans les actes de *Empirical Methods in Natural Language Processing (EMNLP)*.

(Roark, 2002) B. Roark, 2002. Markov parsing : lattice rescoring with a statistical parser. Dans les actes de *Proceedings of the 40th ACL meeting, Philadelphia*.

(Ron et al., 1998) D. Ron, Y. Singer, et N. Tishby, 1998. On the learnability and usage of acyclic probabilistic finite automata. *Journal of Computer and System Sciences 56*(2), 133–152.

(Sadek et al., 1997) D. Sadek, P. Bretier, et F. Panaget, 1997. ARTIMIS : Natural dialogue meets rational agency. Dans les actes de *Fifteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 1030–1035.

(Sadek et al., 1996) M. Sadek, A. Ferrieux, A. Cozannet, P. Bretier, F. Panaget, et J. Simonin, 1996. Effective human-computer cooperative spoken dialogue : The ags demonstrator. Dans les actes de *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, Philadelphia, PA, USA.

(Salmon-Alt et Romary, 2004) S. Salmon-Alt et L. Romary, 2004. Towards a reference annotation framework. Dans les actes de *Proceedings of LREC 2004*.

(Samuelsson et Voutilainen, 1997) C. Samuelsson et A. Voutilainen, 1997. Comparing a linguistic and a stochastic tagger. Dans les actes de *35th conference on Association for Computational Linguistics*, 246–253. Association for Computational Linguistics Morristown, NJ, USA.

(Schapire et Singer, 2000) R. E. Schapire et Y. Singer, 2000. BoosTexter : A boosting-based system for text categorization. *Machine Learning 39*, 135–168.

(Schirra et Joerg, 1993) D. Schirra et R. Joerg, 1993. Connecting Visual and Verbal Space : Preliminary Considerations Concerning the Concept'Mental Image'. Dans les actes de *4th European Workshop Semantics of Time, Space and Movement and Spatio-Temporal Reasoning*.

(Sekine et al., 2002) S. Sekine, K. Sudo, et C. Nobata, 2002. Extended Named Entity Hierarchy. Dans les actes de *Internation Conference on Language Ressources and Evaluation*, 1818–1824.

(Seneff, 1992a) S. Seneff, 1992a. A relaxation method for understanding spontaneous speech utterances. Dans les actes de *Proceedings of the workshop on Speech and Natural Language*, 299–304. Association for Computational Linguistics Morristown, NJ, USA.

(Seneff, 1992b) S. Seneff, 1992b. TINA : A natural language system for spoken language applications. *Computational Linguistics 18*(1), 61–86.

(Servan et Bechet, 2006) C. Servan et F. Bechet, 2006. Décodage conceptuel et apprentissage automatique : application au corpus de dialogue homme-machine media. Dans les actes de *conference annuelle de l'ATALA - Traitement Automatique des Langues Naturelles - TALN*, Leuven.

(Servan et al., 2006) C. Servan, C. Raymond, F. Bechet, et P. Nocera, 2006. Conceptual decoding from word lattices : Application to the spoken dialogue corpus MEDIA. Dans les actes de *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, Pittsburgh, PA, USA, 1614–1617.

(Shannon, 1948) C. Shannon, 1948. A mathematical theory of communcation. *Bell Systems Technical Journal*.

(Shriberg et Stolcke, 2004) E. Shriberg et A. Stolcke, 2004. Prosody modeling for automatic speech recognition and understanding. *Mathematical Foundations of Speech and Language Processing*, 105–114.

(Shriberg et al., 2000) E. Shriberg, A. Stolcke, D. Hakkani-Tur, et G. Tur, 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication 32*(1-2), 127–154.

(Stolcke et Omohundro, 1994) A. Stolcke et S. Omohundro, 1994. Inducing probabilistic grammars by bayesian model merging. Dans les actes de *International Conference on Grammatical Inference*.

(Vapnik, 2000) V. Vapnik, 2000. *The Nature of Statistical Learning Theory*. Springer.

(Vidal et al., 1995) E. Vidal, P. Casacuberta, et P. Garca, 1995. Grammatical inference and applications to automatic speech recognition. *NATO ASI, Speech Recognition and Coding, New Advances and Trends*, 174–191.

(Vidal et al., 1993) E. Vidal, R. Pieraccini, et E. Levin, 1993. Learning associations between grammars : a new approach to natural language understanding. Dans les actes de *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Berlin, Germany.

(Vieira et Poesio, 2000) R. Vieira et M. Poesio, 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics 26(4)*.

(Viterbi, 1967) A. Viterbi, 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on 13*(2), 260–269.

(Walker et al., 2001) M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, et al., 2001. DARPA Communicator Dialog Travel Planning Systems : The June 2000 Data Collection. Dans les actes de *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 1371–1374.

(Walker et al., 1997) M. Walker, D. Litman, C. Kamm, et A. Abella, 1997. PARADISE : A Framework for Evaluating Spoken Dialogue Agents. Dans les actes de *Eighth conference on European chapter of the Association for Computational Linguistics*, 271–280. Association for Computational Linguistics Morristown, NJ, USA.

(Wang et Acero, 2006) Y. Wang et A. Acero, 2006. Rapid development of spoken language understanding grammars. *Speech Communication 48*(3-4), 390–416.

(Wang et al., 2005) Y. Wang, L. Deng, et A. Acero, 2005. Spoken Language Understanding. *IEEE SIGNAL PROCESSING MAGAZINE 1053*(5888/05).

(Wang et al., 2002) Y.-Y. Wang, A. Acero, C. Chelba, B. Frey, et L. Wong, 2002. Combination of statistical and rule-based approaches for spoken language understanding. Dans les actes de *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, Denver, CO, USA, 609–613.

(Whittaker, 2001) E. W. D. Whittaker, 2001. Temporal adaptation of language models. Dans les actes de *Adaptation Methods for Speech Recognition, ISCA Tutorial and Research Workshop (ITRW)*. LM Adaptation for information retrieval of spoken news/radio programs (i.e. SpeechBot).

(Wiebe et al., 2005) J. Wiebe, T. Wilson, et C. Cardie, 2005. Annotationg expressions of opinions and emotions in language. Dans les actes de *Language Resources and Evaluation (formely Computers and the Humanities)*, Volume 39, 165–210.

(Williams et Young, 2007) J. D. Williams et S. Young, 2007. Partially observable markov decision processes for spoken dialog systems. *Computer, Speech and Language 21*, 393–422.

(Wiren et al., 2007) M. Wiren, R. Eklund, F. Engberg, et J. Westermark, 2007. Experiences of an In-Service Wizard-of-Oz Data Collection for the Deployment of a Call-Routing Application. Dans les actes de *Proceeding of the NAACL-HLT workshop, Bridging the Gap : Academic and Industrial Research in Dialog Technologies*, 56–63. Association for Computational Linguistic.

(Woods, 1975) W. Woods, 1975. *What's in a Link : Foundations for Semantic Networks*. Bolt, Beranek and Newman.

(Woods, 1981) W. Woods, 1981. *Procedural Semantics as a Theory of Meaning*. Bolt Beranik and Newman Inc. ; Dist. NTIS.

(Woods, 1983) W. Woods, 1983. Under What Conditions Can a Machine Attribute Meanings to Symbolics (Panel Discussion). *Proc. IJCAI 83*, 47–48.

(Younger, 1967) D. Younger, 1967. Recognition and parsing of context-free languages in time n 3. *Information and Control 10*(2), 189–208.

(Yvon, 2006) F. Yvon, 2006. *Des apprentis pour le traitement automatique des langues*. Paris : l'Université Pierre et Marie Curie - Paris VI.