Apprentissage automatique et compréhension dans le cadre d'un dialogue homme-machine téléphonique à initiative mixte

Christophe Servan

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

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 166 « Information Structures Systèmes - I2S »
Laboratoire Informatique d'Avignon (EA 4128)

Apprentissage Automatique et Compréhension dans le cadre d'un Dialogue Homme-Machine Téléphonique à Initiative Mixte

par

Christophe Servan

Soutenue publiquement le 10 décembre 2008 devant un jury composé de :

M.	Marc El-Bèze	Professeur, LIA, Avignon	Président du Jury
M.	Kamel Smaïli	Professeur, LORIA, Nancy	Rapporteur
M^{me}	Sophie Rosset	Chargée de Recherche, LIMSI-CNRS, Paris-Orsay	Rapporteur
M^{me}	Géraldine Damnati	Docteur, Orange Labs, Lannion	Examinateur
M.	Yannick Estève	Maître de Conférence, LIUM, Le Mans	Examinateur
M.	Renato De Mori	Professeur, LIA, Avignon	Directeur de thèse
M.	Frédéric Béchet	Maître de Conférences, LIA, Avignon	Co-Directeur de thèse

Laboratoire Informatique d'Avignon

Conseil Général de Vaucluse

Remerciements

Tout d'abord, un grand merci à Renato De Mori et Frédéric Béchet qui ont accepté d'être mes directeurs de thèse. Je les remercie de m'avoir fait confiance et de m'avoir fait profiter de leur expérience respective.

Je me permets de remercier tout particulièrement Frédéric Béchet, mon encadrant direct, de ces quelques mots qui ne sont pas à la mesure de la reconnaissance qu'il mérite ni de l'investissement dont il a fait preuve dans cette thèse ; notamment pour sa patience, ses remarques et ses encouragements (*Alors... ça marche ? !*)

Merci à Renato De Mori : je me souviendrais longtemps, avec plaisir, de nos discussions et de nos débats. Merci encore pour tous les précieux conseils prodigués pendant cette thèse.

Je tiens à remercier Kamel Smaïli et Sophie Rosset qui m'ont fait l'honneur d'être mes rapporteurs ainsi que tous les membres de mon jury de thèse, à savoir Marc El-Bèze, Géraldine Damnati et Yannick Estève. Merci de vos remarques qui ont permis d'améliorer la clarté de ce document.

Je remercie le Conseil Général de Vaucluse d'avoir subventionné cette thèse permettant la réalisation de l'ensemble des travaux présentés dans ce manuscrit.

Merci encore à tous ceux qui m'ont permis de près ou de loin à réaliser ce travail de thèse et plus particulièrement mes parents, Jeaninne et Gérard, Mon frère Thierry et sa compagne Mélina et, enfin, ma fiancée Nathalie, pour leur patience et leur bienveillance. Sans eux rien n'aurait pu être possible.

Un grand merci à tous mes amis, plus particulièrement Noémie et Olivier.

Un remerciement particulier à Ludovic avec qui les discussions m'ont souvent apporté un éclairage différent sur mon travail de recherche et souvent beaucoup motivé.

Merci encore à l'ensemble du LIA de m'avoir accueilli et formé au monde de la recherche.

Un grand merci à l'ensemble du personnel du CERI : Laurianne, Nanou, Lolo, Christian, Krusty, Bob, Christophe, Antho, Gilles, Thierry P., Titi, Zak, Ralph, Jens, Mimi, Audrey, Ti'Fred, MJ, Eric C., Mathieu, Ben, Nico S., Nico F., Alex L., Alex P., Alain, Florian B., Florian V., Florian P., Didi, Fred, Georges, Jeff, Patrice, Philou, Corinne, Juan,

Driss, JP, Yézé, Rachid, Renato, Henry, Marc, Abder, Joceline, Simone, Mireille, Laurence, Dom, Cathy, Val et Gisèle (en espérant n'avoir oublié personne...) pour leur joie, leur bonne humeur et tous ces bons moments passés en leur compagnie.

Résumé

Les systèmes de dialogues oraux Homme-Machine sont des interfaces entre un utilisateur et des services. Ces services sont présents sous plusieurs formes : services bancaires, systèmes de réservations (de billets de train, d'avion), etc. Les systèmes de dialogues intègrent de nombreux modules notamment ceux de reconnaissance de la parole, de compréhension, de gestion du dialogue et de synthèse de la parole. Le module qui concerne la problématique de cette thèse est celui de compréhension de la parole.

La compréhension de la Parole : l'approche intégrée

Le processus de compréhension de la parole est généralement séparé du processus de transcription. Il s'agit, d'abord, de trouver la meilleure hypothèse de reconnaissance puis d'appliquer un processus de compréhension. L'approche proposée dans cette thèse est de conserver l'espace de recherche probabiliste tout au long du processus de compréhension en l'enrichissant à chaque étape. Cette approche a été appliquée lors de la campagne d'évaluation MEDIA. Nous montrons l'intérêt de notre approche par rapport à l'approche classique.

En utilisant différentes sorties du module de RAP sous forme de graphe de mots, nous montrons que les performances du décodage conceptuel se dégradent linéairement en fonction du taux d'erreurs sur les mots (WER). Cependant nous montrons qu'une approche *intégrée*, cherchant conjointement la meilleure séquence de mots et de concepts, donne de meilleurs résultats qu'une approche *séquentielle*.

Dans le souci de valider notre approche, nous menons des expériences sur le corpus MEDIA dans les mêmes conditions d'évaluation que lors de la campagne MEDIA. Il s'agit de produire des interprétations sémantiques à partir des transcriptions sans erreur. Les résultats montrent que les performances atteintes par notre modèle sont au niveau des performances des systèmes ayant participé à la campagne d'évaluation.

L'étude détaillée des résultats obtenus lors de la campagne MEDIA nous permet de montrer la corrélation entre, d'une part, le taux d'erreur d'interprétation et, d'autre part, le taux d'erreur mots de la reconnaissance de la parole, la taille du corpus d'apprentissage, ainsi que l'ajout de connaissance *a priori* aux modèles de compréhension.

Une analyse d'erreurs montre l'intérêt de modifier les probabilités des treillis de

mots avec des *triggers*, un modèle cache ou d'utiliser des règles arbitraires obligeant le passage dans une partie du graphe et s'appliquant sur la présence d'éléments déclencheurs (mots ou concepts) en fonction de l'historique.

L'apprentissage Automatique : la diminution de la taille de l'apprentissage

On présente les méthodes à base de d'apprentissage automatique comme nécessairement plus gourmandes en terme de corpus d'apprentissage. En modifiant la taille du corpus d'apprentissage, on peut mesurer le nombre minimal ainsi que le nombre optimal de dialogues nécessaires à l'apprentissage des modèles de langages conceptuels du système de compréhension. Des travaux de recherche menés dans cette thèse visent à déterminer quel est la quantité de corpus nécessaire à l'apprentissage des modèles de langages conceptuels à partir de laquelle les scores d'évaluation sémantiques stagnent. Une corrélation est établie entre la taille de corpus nécessaire pour l'apprentissage et la taille de corpus afin de valider le guide d'annotations. En effet, il semble, dans notre cas de l'évaluation `MEDIA`, qu'il ait fallu sensiblement le même nombre d'exemple pour, d'une part, valider l'annotation sémantique et, d'autre part, obtenir un modèle stochastique « de qualité » appris sur corpus.

De plus, en ajoutant des données *a priori* à nos modèles stochastiques, nous réduisons de manière significative la taille du corpus d'apprentissage nécessaire pour atteindre les même scores du système entièrement stochastique (près de deux fois moins de corpus à score égal). Cela nous permet de confirmer que l'ajout de règles élémentaires et intuitives (chiffres, nombres, codes postaux, dates) donne des résultats très encourageants. Ce constat a mené à la réalisation d'un système hybride mêlant des modèles à base de corpus et des modèles à base de connaissance.

Application à une tâche de routage d'appel : intégration de la compréhension et réduction des coûts d'apprentissage

Dans un second temps, nous nous appliquons à adapter notre système de compréhension à une application de dialogue simple : un système de routage d'appel. La problématique de cette tâche est le manque de données d'apprentissage spécifiques au domaine. Nous la résolvons en partie en utilisant divers corpus déjà à notre disposition. Lors de ce processus, nous conservons les données génériques acquises lors de la campagne `MEDIA` et nous y intégrons les données spécifiques au domaine.

Nous montrons l'intérêt d'intégrer une tâche de classification d'appel dans un processus de compréhension de la parole spontanée. Malheureusement, nous disposons de très peu de données d'apprentissage relatives au domaine de la tâche. En utilisant notre approche intégrée de décodage conceptuel, conjointement à un processus de filtrage, nous proposons une approche sous forme de *sac de mots et de concepts*. Cette ap-

proche exploitée par un classifieur permet d'obtenir des taux de classification d'appels encourageants sur le corpus de test, alors que le WER est assez élevé. L'application des méthodes développées lors de la campagne MEDIA nous permet d'améliorer la robustesse du processus de routage d'appels.

Abstract

Spoken dialogues systems are interfaces between users and services. Simple examples of serivces for which theses dialogue systems can be used include : banking, booking (hotels, trains, flights), etc. Dialogue systems are composed of a number of modules. The main modules include Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Dialogue Management and Speech Generation. In this thesis, we concentrate on the Spoken Language Understanding component of dialogue systems.

Spoken language understanding : the integrated approach

In the past, it has usual to separate the Spoken Language Understanding process from that of Automatic Speech Recognition. First, the Automatic Speech Recognition process finds the best word hypothesis. Given this hypothesis, we then find the best semantic interpretation. This thesis presents a method for the robust extraction of basic conceptual constituents (or concepts) from an audio message.

The conceptual decoding model proposed follows a stochastic paradigm and is directly integrated into the Automatic Speech Recognition process. This approach allows us to keep the probabilistic search space on sequences of words produced by the Automatic Speech Recognition module, and to project it to a probabilistic search space of sequences of concepts. The experiments carried out on the French spoken dialogue corpus MEDIA, available through ELDA, show that the performance reached by our new approach is better than the traditional sequential approach.

As a starting point for evaluation, the effect that deterioration of word error rate (WER) has on SLU systems is examined though use of different ASR outputs. The SLU performance appears to decrease lineary as a function of ASR word error rate. We show, however, that the proposed integrated method of searching for both words and concets, gives better results to that of a traditionnanl sequential approach.

In order to validate our approach, we conduct experiments on the `MEDIA` corpus in the same assessment conditions used during the `MEDIA` campaign. The goal is to produce error-free semantic interpretations from transcripts. The results show that the performance achieved by our model is as good as the systems involved in the evaluation campaign.

Studies made on the MEDIA corpus show the concept error rate is related to the word error rate, the size of the training corpus and *a priori* knwoledge added to conceptual model languages.

Error analyses show the interest of modifying the probabilities of word lattice with *triggers*, a template cache or by using arbitrary rules requiring passage through a portion of the graph and applying the presence of triggers (words or concepts) based on history.

Machine learning : reducing the size of the training corpus

Methods based on machine learning are generally quite demanding in terms of amount of training data required. By changing the size of the training corpus, the minimum and the optimal number of dialogues needed for training conceptual language models can be measured.

Research conducted in this thesis aims to determine the size of corpus necessary for training conceptual language models from which the semantic evaluation scores stagnated. A correlation is established between the necessary corpus size for learning and the corpus size necessary to validate the manual annotations. In the case of the `MEDIA` evaluation campaign, it took roughly the same number of examples, first to validate the semantic annotations and, secondly, to obtain a "quality" corpus-trained stochastic model.

The addition of *a priori* knowledge to our stochastic models reduce significantly the size of the training corpus needed to achieve the same scores as a fully stochastic system (nearly half the size for the same score). It allows us to confirm that the addition of basic intuitive rules (numbers, zip codes, dates) gives very encouraging results. It leeds us to create a hybrid system combining corpus-based and knowledge-based models.

Application to the call routing task : SLU integration into a simple task and reduction of training costs

The second part of the thesis examines the application of the understanding module to another simple dialogue system task, a callrouting system. A problem with this specific task is a lack of data available for training the requiered language models. We attempt to resolve this issue by supplementing he in-domain data with various other generic corpora already available, and data from the MEDIA campaing.

We show the benefits of integrating a call classification task in a SLU process. Unfortunately, we have very little training corpus in the field under consideration. By using our *integrated* approach to decode concepts, along with an *integrated* process, we propose a *bag of words and concepts* approach. This approach used by a classifier achieved encouraging call classification rates on the test corpus, while the WER was relatively

high. The methods developed are shown to improve the call routing system process robustness.

Table des matières

Chapitre 1

Introduction

Qu'est-ce que la compréhension ? La définition que l'on trouve dans un dictionnaire est la suivante : « faculté de comprendre, intelligence ; possibilité d'être compris, intelligible ». Lorsqu'on dialogue avec un serveur vocal, ce dernier doit pouvoir comprendre et formuler une réponse [De Mori, 1998]. Les serveurs vocaux interactifs sont une avancée technologique que beaucoup de personnes attendent depuis des années. Certainement grâce à ce célèbre film d'anticipation de Stanley Kubrick *2001 : A Space Odyssey*. En dehors de tout aspect scénaristique, nous découvrons un personnage central de l'histoire : un ordinateur. Cet ordinateur, le HAL 9000, a la particularité de pouvoir communiquer en langue naturelle.

À l'époque, ce n'était que de l'anticipation, aujourd'hui nous ne sommes plus en 2001 mais en 2008, où en sommes-nous actuellement ? Avec l'explosion de la téléphonie mobile, nous avons vu apparaître des serveurs à commandes vocales simples pour pouvoir consulter le solde de notre compte en banque, consulter notre répondeur, réserver un billet d'avion, etc. Nous sommes assez loin de l'image de Dave parlant avec HAL. Cependant, ce type de serveur, permettant la communication orale Homme-Machine, n'est pas si utopique. Les systèmes de reconnaissance automatique de la parole, ce que l'on pourrait assimiler comme étant l'oreille d'une machine, ont atteint des performances suffisantes pour permettre l'exploitation des informations contenues dans la parole. Par exemple, les outils de dictée vocale sont au point. Nous définissions le terme de « parole » comme étant l'« usage concret de la langue par un individu de manière orale ». À partir de cette définition nous pouvons assimiler la langue naturelle orale à la parole.

Mais ce qui nous intéresse plus que le contenu de la parole, c'est le sens du message que transmet un humain à une machine. En effet, la compréhension c'est aussi l'« ensemble des caractères, des attributs dans un concept ». Un concept étant une unité sémantique de base. Cette définition restreint le cadre de notre étude, en assimilant la compréhension à un processus d'annotation ou d'étiquetage automatique de la parole de manière sémantique. Ce processus est assimilable à un processus de mise en forme de la parole, comme la mise en forme d'une requête dans une base de donnée. Cependant, ce travail d'annotation conceptuel, qu'il soit manuel ou automatique, ne peut se

faire sans un ensemble structuré de termes, de concepts et de relations représentant le sens d'une information. Cet ensemble est appelé ontologie. Le processus de compréhension de la parole, tel que nous venons de le définir, est au cœur de la problématique de cette thèse.

Nous verrons, dans ce mémoire de thèse, différents moyens ou stratégies permettant cette compréhension dans le contexte d'un dialogue entre un humain et une machine, les particularités de la langue naturelle et les difficultés inhérentes au décodage conceptuel pour poser les premières briques d'une communication Homme-Machine Orale.

1.1 La compréhension dans le contexte d'un dialogue Homme-Machine

Il existe deux grandes familles de systèmes de compréhension de la parole, la première concerne les systèmes à base de connaissances et la seconde regroupe les systèmes à base de corpus. Chacune de ces familles a ses avantages et ses inconvénients.

Les systèmes à base de connaissances nécessitent qu'un être humain intervienne pour écrire l'ensemble des règles de compréhension et que l'homme puisse prévoir tous les cas possibles. Ces systèmes nécessitent une description de l'ontologie très précise qui sera la base de la création des règles. Il est à noter qu'au plus l'ontologie est riche, plus l'écriture des règles est longue et coûteuse. Ils effectuent généralement des analyses syntaxiques avant d'effectuer une analyse sémantique. Ils sont généralement peu permissifs dans le vocabulaire et la grammaire du locuteur et leur domaine sémantique est généralement très restreint. Ce sont des systèmes qui s'adaptent généralement assez bien sur des textes ou des paroles structurés, telles que les informations et articles de journaux.

Les systèmes à base de corpus ne disposent pas nécessairement de connaissances *a priori*, c'est pourquoi ils nécessitent beaucoup de données d'apprentissage. Mais l'annotation sémantique d'un corpus est particulièrement longue et coûteuse. Ils sont très souples et permissifs dans le décodage d'une phrase. En effet, même si le système n'a jamais rencontré la séquence de mots présente, il arrivera à produire une solution, celle la plus probable. Ceci est vrai uniquement si les mots, de la séquence de mots analysée, sont connus du système. Cette souplesse est appréciée lors du décodage conceptuel de la parole spontanée. En effet, chaque locuteur étant différent, il s'exprime de manière différente, avec des structures de phrases différentes. Les systèmes à base de corpus sont très permissifs dans la syntaxe et le vocabulaire du locuteur et permettent d'implémenter plus facilement des domaines sémantiques vastes, que les systèmes à base de connaissances.

Il existe cependant une troisième famille de systèmes : les systèmes « mixtes » ou « hybrides ». Ils utilisent à la fois des corpus et des connaissances *a priori* comme, par exemple, les systèmes Chronus et HUM [Pieraccini *et al.*, 1991, Miller *et al.*, 1994].

Tous ces systèmes ne font qu'une chose : représenter le message exprimé par un

locuteur dans un format interprétable par la machine. Ce format est appelé la *représentation sémantique de la langue*. Cette représentation peut revêtir plusieurs formes, mais celle qui est privilégiée est celle sous forme de « boîtes », telle que dans le projet FrameNet [Baker *et al.*, 1998].

La campagne d'évaluation TECHNOLANGUE/EVALDA/MEDIA a permis l'évaluation des performances des systèmes de compréhension pour le français, dans le cadre d'une réservation d'hôtel. Cette campagne a permis de mettre au point une ontologie, un guide d'annotation, un corpus d'apprentissage, un corpus de développement et un corpus de test. Avoir participé à cette campagne d'évaluation, nous a permis de nous comparer à différents systèmes utilisant différentes approches et poser une base qui est actuellement utilisée dans le cadre du projet européen LUNA [1].

Enfin, les systèmes de compréhension de la parole sont utilisés dans de nombreux autres domaines, notamment celui du routage d'appels, qui permet, par un traitement sémantique, de ne conserver que les données porteuses de sens et utiles pour le système.

1.2 Travaux réalisés

L'ensemble de ma thèse a été financée grâce à une bourse octroyée par le Conseil Général de Vaucluse.

La campagne d'évaluation sur les systèmes de compréhension TECHNOLANGUE/EVALDA/MEDIA nous laisse à disposition l'intégralité de son corpus. Grâce à ces données, nous pouvons développer un système complet de dialogue en langue naturelle en français. Une grande partie du travail de thèse réalisé est associé à cette campagne d'évaluation, pour laquelle j'ai réalisé le système de compréhension de l'équipe de recherche, dont je fais partie.

Avec pour objectif d'obtenir le meilleur système de compréhension possible, notre travail est de mettre au point un système de compréhension de la parole. Tout l'intérêt de concevoir un système ayant la possibilité de s'adapter à la langue naturelle est de pouvoir corriger les éventuelles erreurs du module de reconnaissance de la parole (RAP). En effet, les systèmes de RAP sont victimes des différentes difficultés de langage ou des conditions d'enregistrement. Ces problèmes provoquent des erreurs de reconnaissance, qui peuvent être corrigées grâce à un module de compréhension [Bangalore et Johnston, 2004].

Le travail de recherche consiste à rechercher la meilleure stratégie de décodage, permettant d'utiliser indifféremment une séquence de mot ou un treillis de mots. L'approche proposée devra tenir compte du treillis de mots et permettre de décoder ce treillis pour pouvoir, en sortie, choisir le meilleur chemin ou utiliser un gestionnaire de dialogue qui ira chercher les informations qui l'intéressent, dans ce treillis d'hypothèses.

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

La seconde partie du travail de recherche a été d'appliquer le système de compréhension développé pour la campagne MEDIA à une autre tâche de dialogue plus simple : la caractérisation d'appel. Le système de compréhension sera intégré à cette tâche tel que dans l'application *How May I Help You ?* [Gorin *et al.*, 1997]. Nous étudierons la possibilité d'utiliser un tel système malgré le manque de données et nous observerons les performances.

1.3 Organisation du mémoire de thèse

Ce mémoire de thèse contient quatre parties. Tout d'abord, une description du domaine : le dialogue Homme-Machine. Nous verrons alors ce qu'est un système de dialogue et quels sont les différents systèmes existants. Ensuite, nous étudierons les systèmes de compréhension, qui sont une partie des systèmes de dialogue. Dans cette partie, sera exposé l'état actuel des connaissances de ce domaine, lequel est le fil conducteur de ce travail de thèse. Nous présenterons alors les principes des différentes approches inhérentes à la compréhension d'un message oral.

Les parties suivantes (chapitres 3 et 4) sont consacrées au cœur de ma thèse : la mise au point d'un système de compréhension de la parole, pour la campagne d'évaluation MEDIA. Le travail présenté et les résultats obtenus, dans la partie expérience (chapitre 4), découlent de cette campagne et approfondissent une approche permettant de lier les modules de compréhension et de Reconnaissance Automatique de la Parole (RAP). Cette opération a pour but de prendre la décision de la meilleure interprétation le plus tard possible. C'est à dire, à la sortie de ces deux modules et non de manière séquentielle comme il est fait traditionnellement. Nous observerons par le biais d'expériences, l'influence du taux d'erreur mot de la RAP sur la compréhension. Cette partie est conclue par des expériences faites sur l'historique de dialogue.

Dans la quatrième partie de ce mémoire (chapitre 5), il est présenté un système de dialogue Homme-Machine simple : un système de routage d'appel. L'intérêt de cette application est de mettre en évidence la possibilité de réduire le coût de l'apprentissage des modèles de langage, en réutilisant des corpus déjà existants. Ce travail permet de montrer la complémentarité des corpus choisis pour cette tâche. Nous montrerons aussi la flexibilité du système mis au point dans la partie précédente, en utilisant une nouvelle ontologie pour l'annotation sémantique. Enfin, par une approche basée sur un processus de classification automatique, nous présenterons les différentes expériences concernant la catégorisation et le routage d'appels.

Dans la dernière partie, il est proposé quelques conclusions et perspectives relatives aux travaux réalisés et présentés dans ce mémoire.

Chapitre 2

Dialogue Homme-Machine et compréhension de la parole

Sommaire

2.1 Les systèmes de Dialogue

La définition que l'on peut trouver dans un dictionnaire du mot *dialogue* est « conversation, échange de propos entre deux ou plusieurs personnes ». Il existe aussi une définition du *dialogue Homme-Machine* : « échange interactif entre un utilisateur et un ordinateur ». De nos jour, il est de plus en plus fréquent de communiquer avec des machines. La plupart des systèmes de dialogue rencontrés au quotidien permettent à un humain d'entrer en relation avec un service, par exemple, un système de réservation d'hôtel.

Les deux exemples de dialogue Homme-Machine 2.1 et 2.2 sont typiques de ceux que l'on peut trouver actuellement et qui sont largement répandus. À l'inverse, l'exemple de dialogue 2.3 est en langue naturelle. Les trois exemples sont de difficulté croissante pour la machine et sont basés sur le même scénario : la réservation d'un hôtel pour deux personnes à Paris le 23 décembre.

Machine :	Bonjour et bienvenue dans notre service de réservation d'hôtel. Pour réserver un hôtel, tapez 1 ; modifier ou annuler votre réservation tapez 2 ; sinon merci de raccrocher.
Utilisateur :	*Appuie sur la touche 1*
Machine :	Veuillez entrer le code postal de la ville désirée.
Utilisateur :	*Appuie successivement sur les touche 7, 5, 0, 0 et 0*
Machine :	Veuillez entrer le nombre de personnes.
Utilisateur :	*Appuie sur la touche 2*
Machine :	Veuillez entrer successivement le jour du mois puis le mois de la date de la réservation.
Utilisateur :	*Appuie sur la touche 2, 3, 1, 2*
...	

TAB. 2.1: *Exemple de dialogue par touches digitales*

Nous pouvons voir dans le premier exemple 2.1, que le dialogue est très guidé et qu'il ne peut y avoir d'ambiguïtés dans les choix du locuteur. En ce qui concerne la reconnaissance, il n'y pas de difficultés à reconnaître les impulsions sonores des touches digitales. En revanche, le déroulement du dialogue est long, étant donné qu'on ne peut donner qu'une seule information à chaque tour de parole. De plus, le locuteur ne peut que réagir par rapport aux questions de la machine.

Dans le second exemple (encadré 2.2), nous avons une ouverture du dialogue. En effet, le locuteur prononce des mots-clés permettant de valider les choix proposés par la machine. Les probabilités qu'il puisse y avoir des ambiguïtés est faible. Le module de

Machine :	Bonjour et bienvenue dans notre service de réservation d'hôtel. Pour réserver un hôtel dites « hôtel » ; modifier ou annuler votre réservation dites « modifier » ou bien « annuler» ; sinon merci de raccrocher.
Utilisateur :	hôtel
Machine :	Veuillez prononcer la ville désirée.
Utilisateur :	paris
Machine :	Veuillez prononcer le nombre de personnes.
Utilisateur :	deux
Machine :	Veuillez prononcer la date telle que pour le premier janvier « un un».
Utilisateur :	vingt-trois douze
...	

TAB. 2.2: *Exemple de dialogue par commande vocales*

reconnaissance de la parole dispose d'un vocabulaire très faible ce qui limite les erreurs de reconnaissance. On retrouve la lourdeur et la rigidité du dialogue pas à pas, comme dans l'exemple précédent.

Machine :	Bonjour et bienvenue dans notre service de réservation d'hôtel. Que puis-je faire pour vous ?
Utilisateur :	Euh... je voudrais réserver un hôtel à Paris pour le 23 décembre prochain.
Machine :	Vous souhaitez effectuer une réservation d'hôtel à Paris le 23 décembre 2008 pour combien de personnes ?
Utilisateur :	Pour deux personnes.
...	

TAB. 2.3: *Exemple de dialogue en langue naturelle*

Le dernier exemple de dialogue (encadré 2.3), est un dialogue libre. Le locuteur ne se contente pas de mots clés mais formule des réponses complètes. Le dialogue est fluide, rapide et ouvert. Ce dernier exemple, est caractéristique des dialogues Homme-Machine en langue naturelle orale. Les travaux présentés dans le cadre de cette thèse sont associés à ce type de dialogue.

Nous pouvons voir à travers ces exemples que la langue naturelle permet de transporter dans un message vocal, plus d'informations dans un seul tour de dialogue qu'au moyen d'autres systèmes existants actuellement. Mais dans les deux exemples, il n'y a pas d'ambiguïté possible, ni d'expressions implicites. Au contraire, la langue naturel peut être implicite et ambigu, ce qui nécessite un traitement particulier.

2.1.1 Archtecture d'un système de dialogue Homme-Machine

Un système de dialogue oral en langue naturelle est une interface Homme-Machine, qui permet à un utilisateur d'interagir avec une machine. La difficulté vient du fait que la machine doit réagir tel qu'un humain aurait agi à sa place, sans que l'interlocuteur Humain n'en soit gêné. Le système doit comprendre le sens des paroles prononcées par

l'utilisateur, interpréter ce sens, rechercher, puis, trouver une réponse et enfin, formuler cette réponse de manière à répondre aux attentes de l'interlocuteur.

Il existe plusieurs étapes dans un dialogue Homme-Machine :

- la reconnaissance de la parole : cette première partie est celle qui va transformer le signal de parole en symboles permettant son traitement ;
- le module de compréhension : ce dernier est celui qui, à partir des symboles produits par la reconnaissance, va produire une représentation sémantique interprétable par le système ;
- le gestionnaire de dialogue : il va utiliser les informations de l'interprétation pour la traiter (par exemple, en lançant des requêtes dans une base de données) et formuler la réponse ;
- le synthétiseur de parole : ce dernier module va produire une réponse vocale à partir de celle formulée par le gestionnaire de dialogue.

La figure 2.1 montre l'agencement des différents modules inhérents à un système de dialogue.

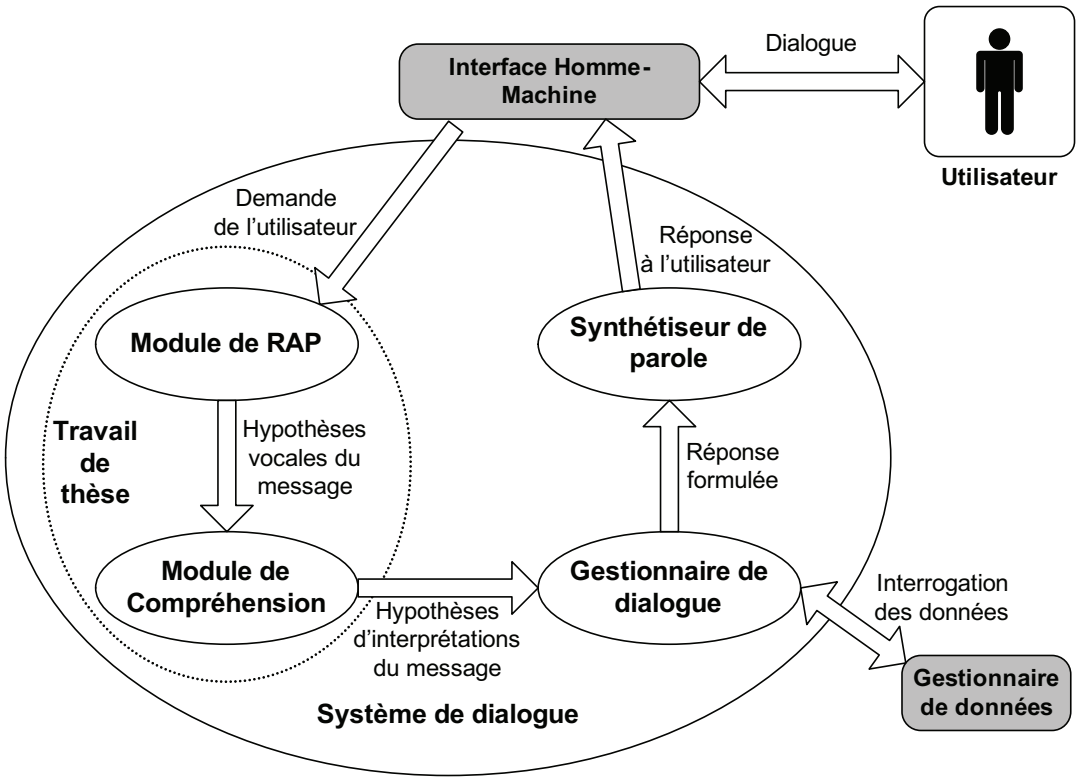

FIG. 2.1: *Exemple de système de dialogue dans lequel on peut voir apparaître les différents modules et leurs combinaisons entre-eux.*

Seuls les deux premiers modules concerneront le travail de thèse présenté dans ce mémoire, ainsi dans la partie 2.2 nous aborderons une rapide présentation de la Reconnaissance Automatique de la Parole. Dans la partie 2.3 sera abordée la compréhension

et en particulier l'interaction possible entre un module de reconnaissance et un module de compréhension.

Pour plus d'informations sur le gestionnaire de dialogue et le synthétiseur de parole, voir [De Mori, 1998].

2.2 Reconnaissance Automatique de la Parole

Le fonctionnement de la grande majorité des systèmes de reconnaissance automatique de la parole (RAP) est fondé sur des principes probabilistes. Ainsi, à partir de la séquence d'observations acoustiques $A = a_1, a_2, ...a_m$ extraite du signal de parole, un système de reconnaissance de la parole recherche la séquence de mots $\hat{W} = w_1, w_2, ...w_k$ qui maximise la probabilité *a posteriori* $P(W|A)$. Cette probabilité n'est pas calculable en l'état car A appartient à l'espace infini. Le théorème de Bayes permet de reformuler cette probabilité en :

$$P(W|A) = \frac{P(W)P(A|W)}{P(A)} \tag{2.1}$$

La séquence de mot $\hat{W}$ est celle qui maximise le produit de l'équation 2.1, comme noté dans l'équation 2.2.

$$\hat{W} = \underset{W}{argmax} P(W|A) = \underset{W}{argmax} \frac{P(W)P(A|W)}{P(A)} \tag{2.2}$$

Or, dans ce produit la probabilité *a priori* de la séquence d'observations acoustiques $P(A)$ n'est pas calculable directement pour les mêmes raisons que $P(W|A)$. Mais ici la séquence d'observations acoustiques A est identique pour toutes les hypothèses W, $P(A)$ est une valeur constante et positive qu'il est inutile de calculer pour trouver $\hat{W}$. On a donc :

$$\hat{W} = \underset{W}{argmax} P(A|W)P(W) \tag{2.3}$$

L'étape de reconnaissance revient à maximiser le produit des 2 termes, $P(A|W)$ et $P(W)$. Le premier terme $P(A|W)$ est la probabilité qu'une séquence de mots W génère la séquence d'observations acoustiques A. Elle est estimée par un modèle acoustique. Le second terme $P(W)$ est la probabilité *a priori* de la séquence de mots, elle est estimée par un modèle de langage.

2.2.1 Modélisation acoustique

La fonction du modèle acoustique est d'estimer le second facteur de l'équation 2.3. Étant donnée une séquence de vecteurs de paramètres extraits du signal de parole, le but des modèles acoustiques est de calculer la probabilité qu'un événement linguistique particulier (un mot, une phrase, etc.) ait généré cette séquence. La plupart des systèmes

actuels recourent à l'utilisation des Modèles de Markov Cachés (*HMMs*). Les unités de base modélisées par ces systèmes sont souvent les phonèmes. L'utilisation de cette unité de base donne de bons résultats de reconnaissance, qui peuvent être améliorés en prenant en compte certaines caractéristiques des signaux de paroles. Il est reconnu que la réalisation d'un phonème est fortement influencée par les phonèmes qui l'entourent (*cf.* chapitre 2 de [Haton *et al.*, 2006]).

Une description plus réaliste des sons de base peut être effectuée en dédiant différents modèles au même phonème selon le contexte. Par exemple, les modèles allophones ou *Tri*-phones. Un modèle de Markov peut être vu comme un automate stochastique. Les états sont associés à une densité de probabilité qui modélise les formes rencontrées $P(Xk|Ei)$ qui est la probabilité de produire l'événement Xk sur l'état Ei. Les transitions assurent les contraintes d'ordre temporel, des formes pouvant être observées. $P(Ei+1|Ei)$ est la probabilité de transition d'un état à un autre.

Les HMMs s'appuient sur un corpus d'apprentissage des différentes réalisations des phonèmes de la langue considérée pour les modéliser. Un modèle de mot peut alors être vu comme une concaténation successive de modèles de phonèmes, la phrase complète comme une concaténation de modèles de mots. Les chemins, dans cet automate, représentent toutes les chaînes possibles de mots. Le chemin ayant la plus forte probabilité est donc celui qui permet l'alignement optimal du signal acoustique sur le modèle de Markov. L'opération appelée décodage est celle qui permet de retrouver au mieux ce meilleur chemin. Différents algorithmes existent : Viterbi, le Beam Search et le A*.

2.2.2 Modélisation statistique de la langue

De façon générale, on appelle langue un ensemble de conventions permettant de conceptualiser la pensée et d'échanger avec d'autres interlocuteurs à l'aide d'un support comme la parole, l'écriture, etc. L'objectif de la modélisation de la langue est de trouver un moyen de le décrire.

Deux grandes approches existent : d'une part, la modélisation basée sur l'utilisation de grammaires formelles mises au point par des experts en linguistique et, d'autre part, la modélisation stochastique qui tente de décrire automatiquement un langage à partir de l'observation de corpus. Il est long et très difficile de représenter un langage naturel de manière formelle. Dans le cadre d'applications de dialogue oral autorisant l'utilisateur à s'exprimer librement, la modélisation stochastique de la langue est souvent privilégiée. Cependant, on trouve beaucoup plus de systèmes à base de règles formelles, la plupart des systèmes existant restent dans un domaine d'utilisation restreint.

Les modèles de langage statistiques ont pour objectif d'estimer la seconde composante de l'équation 2.3. En effet, le modèle acoustique n'a pas de connaissances *a priori* sur l'enchaînement entre les mots. Il est mentionné dans [Mariani, 1990] qu'une suite de 9 phonèmes peut être transcrite en 32 000 suites de mots différentes, dont quelques unes seulement sont syntaxiquement correctes. Les modèles de langage vont permettre de représenter des connaissances d'ordre linguistique afin de guider le décodage vers des hypothèses de phrase cohérentes du point de vue syntaxique ou grammatical. Les

modèles de langage statistiques tentent alors de déterminer la probabilité *a priori* de la séquence de mots $W_k = w_1, w_2, ..., w_k$ selon l'équation 2.4.

$$\begin{aligned} P(W_k) &= \prod_{i=1}^{k} P(wi|hi) \\ où\ h_i &= w_1, ..., w_{i-1}\ pour\ i > 2 \\ h_i &= w_1\ pour\ i = 2 \\ h_i &= ''\ ''\ pour\ i = 1 \end{aligned} \tag{2.4}$$

Le principal problème dans l'utilisation de modèles de langage probabilistes tient en la longueur de l'historique considérée. En effet, la taille des corpus d'apprentissage ne nous permet pas de calculer efficacement la probabilité $P(w_i|w_1, w_2, ..., w_{i-1})$. On approche alors la probabilité en fonction d'un historique de taille réduite et fixe. C'est ce que l'on nomme un modèle N-grammes. Le calcul considère alors, pour la prédiction d'un mot, que la suite des $N-1$ mots qui le précèdent est suffisante. Un N trop petit modélise mal les contraintes linguistiques tandis qu'un N trop grand va grandement limiter la couverture du modèle. Les valeurs les plus utilisées sont $N = 2$ pour des modèles de langage appelé *Bi*-grammes et $N = 3$ pour des modèles de langage appelés *Tri*-grammes. Les termes de l'équation 2.4 pour un modèle *Tri*-gramme se résument alors à l'équation 2.5.

$$P(W_k) = P(w_1) \times P(w_2|w_1) \times \prod_{i=3}^{k} P(w_i|w_{i-2}w_{i-1}) \tag{2.5}$$

La probabilité d'apparition d'un mot est généralement estimée par le critère de maximum de vraisemblance. Pour un modèle *Tri*-gramme et pour un mot w précédé des mots $w_i w_j$ cela donne :

$$P(w|w_i w_j) = \frac{c(w_i w_j w)}{c(w_i w_j)} \tag{2.6}$$

où $c(w_i w_j w)$ correspond au nombre d'occurrences de la suite de mots $w_i w_j w$ dans le corpus d'apprentissage et $c(w_i w_j)$ au nombre d'occurrences de la suite de mots $w_i w_j$. Mais les événements non-observés dans le corpus d'apprentissage du modèle ont une probabilité nulle avec l'équation 2.6 ; ceci est le problème de cette modélisation. Afin d'y remédier, plusieurs approches ont été développées pour pouvoir modéliser les événements qui n'ont pas été rencontrés lors de la phase d'apprentissage. Certaines utilisent des connaissances sur le langage pour générer des événements manquants, comme les modèles à base de classes (section 2.2.3), d'autres sont des techniques de lissage (section 2.2.4) dont les plus connues sont basées sur des méthodes de repli (ou *back-off*) sur des modèles N-grammes d'ordre inférieur.

2.2.3 Modèle *N*-gramme à base de classes

La quantité de données nécessaires à l'apprentissage d'un modèle de langage robuste et performant, malgré l'utilisation de l'approximation N-grammes, reste impor-

tante. En partant du constat que certains mots ont un comportement similaire, il est possible de les regrouper en classes [Smaïli *et al.*, 1999] :

- le nombre d'événements à modéliser est moindre, il nécessite donc moins de données d'apprentissage ;
- l'utilisation des classes permet d'établir une généralisation : certains événements non-observés au niveau des mots dans le corpus d'apprentissage peuvent être modélisés au niveau des classes.

Dans ce cadre là, un mot w_i appartient à une classe c_i. Il est à noter qu'un mot peut très bien appartenir à plusieurs classes. Pour des raisons de simplicité, supposons qu'un mot w_i n'appartient qu'à une seule classe c_i, le modèle n-classes peut être construit à partir des $n-1$ classes précédentes :

$$P(w_i|c_{i-n+1}...c_{i-1}) = P(w_i|c_i)P(c_i|c_{i-n+1}...c_{i-1}) \tag{2.7}$$

où $P(w_i|c_i)$ est la probabilité du mot w_i dans la classe c_i et $P(w_i|c_{i-n+1}...c_{i-1})$ est la probabilité de la classe c_i connaissant l'historique des $n-1$ classes précédentes. La probabilité d'une phrase W est donnée par :

$$P(W) = \sum_{c:w_i \in c} \prod_i P(w_i|c_{i-n+1}...c_{i-1}) = \sum_{c:w_i \in c} \prod_i P(w_i|c_i)P(c_i|c_{i-n+1}...c_{i-1}) \tag{2.8}$$

Si les classes ont une intersection vide, c'est à dire qu'à un mot ne correspond qu'une seule classe et que le modèle est de type *Tri*-classes, alors l'équation 2.8 peut être simplifiée en :

$$P(W) = \sum_{c:w_i \in c} \prod_i P(w_i|c_i)P(c_i|c_{i-n+1}...c_{i-1}) \tag{2.9}$$

Si $C(w_i)$ est la fréquence du mot w_i, $C(c_i)$ la fréquence de la classe c_i et $C(c_{i-1}, c_i)$ la fréquence qu'un mot d'une classe c_{i-1} soit suivi immédiatement d'un mot de la classe c_i, la probabilité *Bi*-grammes serait :

$$P(w_i|w_{i-1}) \simeq P(w_i|c_{i-1}) = P(w_i|c_i)P(c_i|c_{i-1}) = \frac{C(c_i, w_i)C(c_{i-1}, c_i)}{C(c_i)C(c_{i-1})} \tag{2.10}$$

2.2.4 Lissage

Comme déjà dit, la couverture des corpus d'apprentissage est un problème importants en modélisation stochastique de la langue. Malheureusement, ils ne couvrent pas toutes les successions de mots possibles . Ceci est encore plus vrai dans les applications de dialogue où les corpus sont de taille restreinte. De nombreux événements, des successions de mots possibles, ne sont pas observés. La probabilité qui leur est associée est alors nulle. Une chaîne de mots où apparaît un de ces événements n'est pas considérée comme une transcription potentielle, et ce, sans considération de son score acoustique.

Le but du lissage est de prévenir ces potentielles erreurs de reconnaissance en rendant plus uniforme la distribution observée, en attribuant une probabilité non nulle

à ces événements et en ajustant à la baisse les probabilités trop fortes. Les principales techniques sont détaillées dans [Chen et Goodman, 1996] où est également présentée une discussion sur leurs performances respectives.

Principe

L'estimation des paramètres d'un modèle de langage de type N-grammes est le plus souvent obtenue par la combinaison de deux composants : un modèle de *discounting* (décompte) et un modèle de redistribution. Le principe général est de prélever une quantité à la masse des probabilités issue des événements observés et de la redistribuer aux probabilités associées aux événements non vus.

La probabilité d'un mot jamais vu en présence d'un historique donné est nulle sans lissage. Les méthodes de lissage ici présentées lui attribuent une valeur non nulle calculée à partir d'un historique réduit.

Décompte

La fréquence conditionnelle relative fr d'un mot w selon un historique h s'écrit :

$$\begin{cases} fr(w|h) = \dfrac{c(hw)}{c(h)} & si\ c(h) > 0 \\ fr(w|h) = 0 & si\ c(h) = 0 \end{cases} \tag{2.11}$$

Toutes les méthodes de *discounting* introduisent une fréquence conditionnelle décomptée $fr^*(w|h)$ telle que :

$$0 \leq fr^*(w|h) \leq fr(w|h)\ \forall\ hw \in V^n \tag{2.12}$$

Redistribution

Pour un historique h donné, la redistribution de la masse des probabilités ôtée de fr s'effectue à l'aide d'une composante appelée la probabilité de fréquence nulle (*zero frequency probability*), calculée à partir de fr^*. La probabilité de fréquence nulle, notée l, est définie comme suit :

$$\lambda(h) = 1 - \sum_{w \in V} fr^*(w|h) \tag{2.13}$$

Cette définition implique que, pour un historique jamais observé ($c(h) = 0$), alors $\lambda(h) = 1$. Pour un mot w jamais rencontré après l'historique h, la probabilité de fréquence nulle associée à h est utilisée pour pondérer la valeur de $P(w|h')$, où h' est un historique moins restrictif que h et pour lequel on suppose que l'événement $h'w$ a plus de chance d'avoir été observé que hw.

Lissage par repli (*backing-off*)

Le lissage par repli [Katz, 1987] est un lissage de type hiérarchique. Le principe de cette technique consiste à utiliser un modèle de langage plus général lorsqu'un modèle spécifique ne détient pas suffisamment d'information pour un contexte donné. Par exemple, lorsque pour un n-gramme hw, où h correspond aux $n-1$ mots précédant le mot w, aucune observation n'a été obtenue sur le corpus d'apprentissage, le modèle n-gramme se tourne vers un modèle de niveau inférieur *(n-1)*-gramme : ce processus peut bien sûr être réitéré jusqu'au niveau le plus bas, le *zéro*-gramme, qui consiste en l'attribution d'une constante indépendante du mot w. La probabilité d'un n-gramme est donc estimée à partir du lissage de l'approximation la plus significative (du point de vue de la quantité d'observations) :

$$P(w|h) = \begin{cases} fr^*(w|h) & sifr^*(w|h) > 0 \\ \alpha_h \lambda(h) P(w|h') & sinon \end{cases} \tag{2.14}$$

avec

$$\alpha_h = \left(\sum_{w:fr^*(w|h)=0)} P(w|h') \right)^{-1} \tag{2.15}$$

qui permet à la distribution P(w | h) de respecter la contrainte de sommation à 1.

2.2.5 Combinaison des modèles acoustiques et des modèles de langage

Facteur d'ajustement

Bien que la formule 2.3 suggère que la probabilité du modèle acoustique et la probabilité du modèle de langage peuvent être combinées à travers une simple multiplication, il est nécessaire, en pratique, d'effectuer une pondération. Sans cela, la participation d'un des modèles est négligeable à cause de la différence de l'ordre de grandeur de leurs probabilités. En effet, les probabilités du modèle acoustique sont beaucoup plus petites que celles du modèle de langage : $P(A|W) \ll P(W)$.

La solution la plus couramment utilisée pour atténuer ce problème consiste à ajouter un poids, noté lw (pour *linguistic weight*) et souvent appelé *facteur d'ajustement* (ou *fudge factor*), au modèle de langage. On a alors :

$$\hat{W} = \underset{W}{argmax}\ P(A|W) P(W)^{lw} \tag{2.16}$$

Le poids lw est déterminé empiriquement à partir d'expériences effectuées sur un corpus de développement : la valeur choisie est celle qui optimise les performances du système de reconnaissance. Généralement, $lw > 1$.

Pénalité linguistique

La contribution du modèle de langage peut aussi être interprétée comme une pénalité sur le nombre de mots. En fonction des valeurs des probabilités du modèle de langage, le système peut privilégier une séquence composée de peu de mots longs ou, au contraire, une séquence constituée de nombreux mots courts. Afin d'ajuster au mieux la tendance du système à insérer ou supprimer des mots, une valeur appelée pénalité linguistique et notée *pl* est insérée dans la formule 2.16, qui devient :

$$\hat{W} = \underset{W}{argmax}\ pl^{N(W)} P(A|W) P(W)^{lw} \tag{2.17}$$

où $N(W)$ est le nombre de mots de la séquence W. Tout comme le *facteur d'ajustement* lw, la pénalité linguistique *pl* est déterminée empiriquement : la valeur choisie doit optimiser les performances du système de reconnaissance pour des expériences effectuées sur un corpus de développement.

Utilisation des logarithmes

Les multiplications successives de probabilités, c'est-à-dire de valeurs comprises entre 0 et 1, conduisent à manipuler des valeurs de plus en plus proches de 0. La limite de capacité de représentation de valeurs proches de 0 d'un ordinateur est rapidement atteinte, à moins de mettre en place des mécanismes coûteux en terme de temps de calcul. En pratique, les systèmes de reconnaissance de la parole ne manipulent pas directement les probabilités : ce sont les logarithmes de ces probabilités qui sont utilisés. Le passage aux logarithmes entraîne l'utilisation d'additions plutôt que de multiplications : ce type d'opérations utilise la propriété intéressante des logarithmes qui permet de mettre en évidence les différences entre deux probabilités de manière plus nette. Ainsi, la formule 2.17 se réécrit :

$$\hat{W} = \underset{W}{argmax}(N(W)\log(pl) + \log P(A|W) + lw.\log P(W)) \tag{2.18}$$

2.2.6 Espace de recherche et graphe de mots

À partir de l'observation d'événements acoustiques et de connaissances *a priori* (lexique, modèles acoustiques...), un système de reconnaissance génère un ensemble d'hypothèses de séquences de mots. Cet ensemble est appelé *espace de recherche* : le système doit en extraire la phrase qui satisfait l'équation 2.17. L'*espace de recherche* est généralement représenté sous la forme d'un graphe, appelé graphe de recherche, qui intègre les informations utilisées pour la génération des hypothèses : informations temporelles, unités acoustiques (phonèmes, syllabes, demi-syllabes, ...) associées à leurs scores acoustiques (probabilités données par le modèle acoustique), mots induits par les séquences d'unités acoustiques...

La recherche de la phrase de probabilité maximale au sein d'un graphe de recherche est analogue au problème de la recherche du chemin de poids minimal dans un graphe. De nombreux algorithmes existent pour résoudre ce problème [Cettolo *et al.*, 1998]. Cependant, pour la majorité des systèmes, la taille de l'espace de recherche est très importante et ralentit considérablement le traitement. Pour obtenir une solution dans un délai acceptable, une recherche en faisceau, appelée *beam search*, permet de restreindre l'espace de recherche en supprimant des hypothèses qui semblent localement peu probables [Ney *et al.*, 1992]. Cet élagage ne garantit pas l'obtention de la phrase la plus probable, mais le compromis entre la durée du traitement et la perte de précision est très souvent largement acceptable.

L'utilisation de modèles de langage sophistiqués, par exemple un modèle N-grammes avec un N assez grand ou dans le cas d'un modèles à N variable comme les *multi*-grammes [Zitouni *et al.*, 1999], multiplie les paramètres. La solution la plus répandue consiste à utiliser ce type de modèle lors d'une deuxième passe : le graphe de recherche généré lors d'une première passe est élagué grâce à l'application d'un algorithme de *beam search* et n'est plus composé que de mots. Chaque mot est alors associé à un score acoustique calculé à partir des scores des unités acoustiques qui le composent. Le graphe obtenu pour la deuxième passe est un graphe de mots : il est l'objet de traitements linguistiques lourds qui auraient fortement ralenti le processus de reconnaissance s'ils avaient été appliqués sur l'intégralité de l'espace de recherche dès la première passe.

Généralement, le modèle de langage stochastique utilisé en première passe d'un processus de reconnaissance de la parole est un modèle *Bi*-grammes, voire *Tri*-gramme. Ces modèles ont la particularité d'être simples d'emploi et relativement peu coûteux en temps de calcul. Ces caractéristiques, combinées à l'influence largement bénéfique de ces modèles sur les résultats d'un processus de reconnaissance, sont à l'origine de leur très forte implantation dans les systèmes de reconnaissance de la parole. Les modèles de langage plus évolués, faisant appel à des historiques plus importants ou à des sources de connaissances supplémentaires, sont utilisés en seconde passe sur un espace de recherche réduit à un graphe de mots ou à une liste des N-meilleures hypothèses.

Le graphe de mots ou la liste des N-meilleures hypothèses sont issus du décodage effectué en première passe. Cette seconde passe, qui consiste à utiliser un ou plusieurs modèles nécessitant plus de ressources qu'un modèle *Bi*-grammes afin d'améliorer encore la reconnaissance, est habituellement connue sous le nom de phase de *rescoring*. Bien entendu, rien n'empêche d'utiliser ces modèles de langage gourmands en ressources dans un système de reconnaissance basé sur une seule passe. Malheureusement, les algorithmes utilisés et la technologie actuelle ne permettent pas d'obtenir des résultats satisfaisants dans des délais raisonnables. Dans une application conviviale de dialogue entre un homme et une machine, le système de reconnaissance de la parole doit avoir un temps de réponse proche du temps réel. L'utilisation de systèmes multi-passes permet d'utiliser des modèles de langage nécessitant de grosses ressources sans trop ralentir le processus global de reconnaissance.

2.3 La compréhension de la parole spontanée

2.3.1 Introduction

La sémantique est liée à l'organisation de la signification et à l'ensemble des relations entre les signes sensoriels (ou symboles) et ce qu'ils signifient [W.A.Wood, 1975]. L'interprétation sémantique faite par un ordinateur est un processus de conceptualisation du monde en utilisant l'outil informatique pour créer une structure de représentation de la signification, à partir de différents symboles et de leurs caractéristiques présents dans des mots ou des phrases. La compréhension de la parole spontanée (ou *Speech Language Understanding*) est l'interprétation des symboles transportés par un signal de parole. La compréhension de la parole spontanée et la compréhension de la langue naturelle (ou *Natural Language Understanding*) ont en commun le but d'obtenir une représentation conceptuelle de phrases exprimées dans le langage naturel. La spécification de la compréhension de la parole spontanée est d'avoir à interpréter des symboles contenus dans le signal avec des informations comme l'identification du locuteur. De plus, les phrases prononcées ne suivent pas forcément la structure grammaticale d'une langue. Ces phrases présentent des autocorrections, des hésitations, des répétitions, et d'autres phénomènes acoustiques irréguliers. La compréhension de la parole utilise un module de Reconnaissance Automatique de la Parole (*RAP*) et doit être robuste au bruit engendré par la nature spontanée de la parole et aux erreurs introduites pas la *RAP*. De plus, la sortie du module de *RAP* est un flux de mots sans structure de phrase, ni ponctuation et encore moins de frontières. La compréhension de la parole ne peut donc se fier à ces marqueurs et doit produire une segmentation du texte et un décodage conceptuel en même temps.

Extraire un sens à partir de la parole est un processus complexe ; beaucoup d'approches différentes et de modèles ont déjà été proposés. Les systèmes développés dans les années 70-80 sont basés sur une analyse syntaxique à partir de la meilleure séquence de mots, proposée par un module *RAP*. Ces systèmes utilisent des règles non probabilistes pour faire correspondre des structures syntaxiques avec des structures sémantiques. Ces règles étaient exprimées sous forme d'équations logiques. Nous pouvons trouver dans [W.A.Wood, 1975] une réflexion intéressante concernant les structures des représentations sémantiques pendant cette période. Cet article propose une discussion sur la représentation du sens et les approches afin d'obtenir une de ces représentations sémantiques à partir des mots. Les problèmes élémentaires associés sont listés dans l'article [McTear, 2006]. Dans les années 90, le besoin s'est fait sentir de tester les processus de compréhension sur de grands corpus, ce qui permit l'évaluation automatique de certains paramètres des modèles. Les grammaires et modèles d'interprétation probabilistes sous forme de machine à état finis (ou *Finite State Machines - FSM*) ont été utilisés pour résoudre le problème des ambiguïtés issues des imprécisions de ces modèles. Les systèmes basés sur ces approches sont présentés dans ce chapitre et reprises dans le chapitre 14 de [De Mori, 1998].

D'autres approches consistent à transformer directement les symboles en éléments sémantiques de base pour les agencer ensuite en structures sémantiques. L'intégration

du module de reconnaissance et de compréhension qui est décrite dans cette thèse (chapitre 3) génère des solutions multiples, qui sont ensuite validées en utilisant des contraintes imposées par le contexte dans laquelle une phrase va être interprétée. Le niveau de complexité demandé pour représenter la signification d'un message de parole dépend principalement de l'application ciblée. Il existe trois domaines d'applications : les systèmes de dialogue, la recherche d'information, la traduction automatique.

Nous ne parlerons pas de traduction automatique, car les modèles de compréhension sont très dépendants de la méthode de traduction utilisée, ce qui est ici hors propos.

Les applications de recherche d'information se concentrent sur la recherche d'information sémantique comme les entités nommées, les thèmes et les opinions. Pour ce faire, une représentation sémantique à plat, comme la paire *attribut/valeur*, est la plus souvent utilisée. Les systèmes de dialogue ont besoin de modèles de compréhension pour implémenter des applications de dialogue qui vont plus loin que le routage d'appel ou la complétion de formulaire.

Le projet européen LUNA [1] définit trois niveaux de complexité pour les applications de dialogue. Le premier niveau inclut le processus de traduction de mots en unités sémantiques de base (génération de concepts sémantiques). Ce niveau de complexité est suffisant pour des applications telles que le routage d'appel et la classification pour affiner les catégories. Le second niveau de compréhension effectue une composition sur les concepts de base pour des applications comme le routage d'appel avec une caractérisation de l'énoncé (ceci améliore la compréhension), les systèmes de questions-réponses et les enquêtes de satisfaction. Au troisième niveau de compréhension, un plus large contexte est pris en compte pour une validation sensible au contexte dans des applications de dialogues complexes et des enquêtes de satisfaction, considérant les énoncés comme des ensembles de sous-énoncés et l'interprétation comme un sous-énoncé contextuellement sensible aux autres. Les différents modèles sémantiques et processus d'interprétations qui sont présentés dans ce chapitre appartiennent à un de ces niveaux d'interprétation. Il en va de même pour les applications et les systèmes associés à ces modèles ou processus.

Les différents travaux présentés dans ce mémoire sont focalisés sur les deux premiers niveaux d'interprétation sémantique.

2.3.2 Intérêt de la compréhension dans un dialogue Homme-Machine

Dans le cadre d'une application de dialogue Homme-Machine, le module de compréhension extrait le sens des propos exprimés par le locuteur. Ce module doit extraire les informations qui seront utiles pour l'application. Ces informations utiles seront utilisées par le module de gestion de dialogue pour interroger une base données et formuler une réponse à une requête [Jamoussi *et al.*, 2005]. Le module de compréhension est alors considéré non pas comme un système d'interrogation de base de données mais comme

[1] www.ist-luna.eu

un système de mise en forme de requête. L'extraction des informations sémantiques utiles peut aussi être vue comme la traduction d'une représentation exprimée par le locuteur, vers une représentation interprétable par le système de dialogue. Cette représentation est appelée représentation sémantique que nous détaillerons dans la partie 2.3.4. Cette traduction doit être robuste afin que la machine ne perde pas d'éléments de la requête à formuler.

2.3.3 Difficultés de la parole spontanée

Comme présenté dans [Bousquet-Vernhettes, 2002], il existe diverses difficultés auxquelles est confronté un système de compréhension de la parole. Nous considérons quatre catégories de difficultés provenant de quatre sources différentes :

- Variabilités de la langue et du comportement intra et inter-locuteur ;
- Les difficultés d'ordre linguistique comme les tournures de styles (références, ambiguïtés…) ;
- Particularités de l'oral spontané propre à la parole comme les hésitations, les autocorrections…
- Les erreurs produites par le module de reconnaissance automatique de la parole (RAP) : Les erreurs émises par le module de RAP comme les insertions, délétions et substitutions d'un mots ou d'une séquence de mots. Ces erreurs sont liées généralement aux performances du module de RAP ou à la présence de mots hors-vocabulaire.

Difficultés d'ordre linguistique

Ces difficultés sont liées à la structure même de la langue, en effet, on retrouve ces difficultés aussi bien à l'écrit qu'à l'oral. Il en existe toute une collection, nous retiendrons cependant les plus courantes :

Les ambiguïtés. Il en existe plusieurs type d'ambiguïtés : lexicales, syntaxiques, sémantiques (polysémie, par exemple) ainsi que les confusions de l'utilisateur.

Les formes de styles. Nous utilisons parfois des formes de styles pour nous exprimer (comme l'ironie, les incises…) qui sont bien plus présentes à l'oral qu'à l'écrit.

Les références. Elles sont des mots ou des absences de mots (sous-entendus) qui font référence à d'autres mots ou séquences de mots. Les références peuvent aussi être faites à des lieux ou à des dates (i.e. aujourd'hui, là-bas…)

Les inférences. Elles sont généralement présentes sous la forme d'un ensemble de connaissances qui permettent de replacer une phrase dans son contexte. Ces connaissances sont généralement liées à la tâche et au domaine de traitement. Par exemple, si on considère un système de réservation d'hôtel, lorsque l'utilisateur demandera une réservation pour une ville, le système comprendra qu'il s'agit d'une réservation d'hôtel qui aura lieu dans cette ville.

Il existe bien d'autres formes de style mais elles sont généralement peu utilisées dans le langage naturel oral (i.e. tropes).

Particularités de la parole spontanée

La parole spontanée dispose de deux caractéristiques qui peuvent poser des difficultés aux systèmes de compréhension de la parole. La première tient dans le caractère spontané de la langue. En effet, le locuteur s'exprimant librement, il peut parfois hésiter, se reprendre, s'auto-corriger, répéter des mots ou des séquences de mots, faire des faux-départs, etc. Tous ces phénomènes sont à prendre en compte lorsqu'on traite de la parole spontanée. Par exemple : « Euh…bonjour je euh…oui je voudrais réserver un hôtel à Paris hum oui c'est ça Paris ».

La seconde difficulté vient du fait que le langage oral n'est pas aussi bien structuré que l'écrit. En effet, l'élocution est particulièrement riche en erreurs de grammaire, de conjugaison, d'accord, etc. [Blanche-Benveniste, 1997]. Par exemple : « ils peuvent pas faire ça » ou encore « j'y vais demain ».

Au vu de ces difficultés dues au caractère spontané de la parole, il paraît difficile de concevoir un système de compréhension de la parole de manière identique à un système de compréhension de l'écrit. Il faut donc tenir compte de ces difficultés lors de la conception du système de compréhension.

Variabilités de la langue naturelle

Lorsqu'on met au point une application de dialogue Homme-Machine, il faut tenir compte du fait que tous les locuteurs n'ont pas la même manière de s'exprimer. En effet, des études ont montré que les locuteurs ont des réactions différentes face à une machine [Spérandio et Letang-Figeac, 1986]. Cette variabilité dans les comportements humains, entraîne une variabilité dans la manière de s'exprimer et donc dans le langage.

Ces variations au niveau de la langue prennent plusieurs formes. D'une part, il existe les variabilités inter-locuteur : chaque personne réagit de manière différente face à une machine. Certains vont utiliser un langage particulier et s'adapter à la machine ; d'autres par contre, ne vont faire aucun effort ni d'élocution ni de prononciation. Généralement on distingue deux types d'utlisateurs : les *novices* et les *experts*. D'après des travaux effectués sur le système ARISE [Lavelle *et al.*, 1998], on peut constater que le passage de l'état de *novice* à *expert* s'effectue après peu d'utilisations. Il a même été observé qu'un locuteur peut devenir *expert* dès la première utilisation.

De plus, chaque locuteur, a son propre vocabulaire et sa propre manière de s'exprimer. Cela pose le problème de la couverture lexicale. Outre le fait que ce problème est commun au système de reconnaissance de la parole, si la machine rencontre un mot qu'elle ne connaît pas, elle ne comprendra pas ce que l'utilisateur veut. Il existe plusieurs stratégies dans ce cas. Soit on peut tout simplement ignorer ce que la machine ne comprend pas, après tout l'utilisateur ne commandera pas des pizzas à un système de réservation d'hôtel, soit on tente malgré tout d'essayer de retrouver ce que le locuteur veut dire. Dans tous les cas, le gestionnaire de dialogue peut prendre le relais et initier une demande de répétition ou de précision.

D'autre part, chaque locuteur, ne se met pas spécialement en condition afin d'utiliser la machine. Le comportement d'un locuteur peut varier au cours de l'utilisation [M.Benzequida *et al.*, 2006]. Il peut s'énerver, se calmer, parler vite ou, au contraire, doucement, suivant son état avant d'entamer le dialogue. Ces variations intra-locuteur vont mettre à mal le système de compréhension, qui pourra ne pas comprendre un locuteur qui fait de l'ironie lorsqu'il s'énerve. Par exemple : « Mais quelle machine performante… »

Enfin, la polysémie, étant déjà source d'ambiguïté, fait aussi partie des variabilités de la langue. En effet, les problèmes de recouvrement sémantique inhérents à des concepts trop proches où à des concepts utilisant les mêmes mots mais, dans un sens différent. Par exemple : « l'hôtel d'Avignon », cette phrase peut aussi bien signifier l'hôtel qui se trouve dans la ville d'Avignon, que le nom d'un hôtel s'appellant « l'Hôtel d'Avignon ».

Erreurs du module de reconnaissance

Dans cette thèse, un des axes principaux des travaux est la compréhension de la parole. Or, pour effectuer ce travail de compréhension, le système nécessite un module de reconnaissance automatique de la parole (RAP). De ce fait, il est évident que les erreurs provoquées par le module de RAP se répercuteront sur le module de compréhension. Par exemple, si le système de RAP ne reconnaît pas le nom d'une ville, le système de compréhension ne pourra l'identifier comme telle. Comme déjà énoncé dans la section précédente, ces erreurs peuvent être dues à un problème de couverture lexicale. En effet, le système de RAP ne peut reconnaître un mot que s'il existe dans son lexique. Dans l'exemple précédent, si le système de RAP ne connaît pas le nom de la ville, il ne pourra jamais le reconnaître.

Enfin il est à noter que les variabilités d'élocution sont aussi sources d'erreurs. Par exemple, le fait de parler trop vite ou d'avoir un accent trop prononcé peut mettre en défaut le système de RAP. Ce qui provoquera des insertions, des substitutions et des délétions que le système de compréhension aura plus de mal à décoder que si ces erreurs n'étaient pas présentes.

Ce problème fait partie de ceux traités dans mes travaux de thèse. En effet, nous avons travaillé à la mise au point d'une stratégie permettant de rectifier, en partie, la reconnaissance proposée par le module de reconnaissance dans le chapitre 3.

2.3.4 Représentation informatique de la sémantique.

La représentation sémantique de la langue (ou Meaning Representation Language - MLR) a sa propre syntaxe. La MLR, ou ontologie, doit suivre un modèle de représentation cohérent avec la théorie sémantique. Elle tient compte des relations de raisonnement et de composition des constituants sémantiques pour les relier entre eux dans

des structures et des procédures. Il est difficile de concevoir une représentation sémantique permettant de décrire l'ensemble des expressions d'un message vocal quel qu'il soit. Afin de produire des systèmes réalisables, la représentation tend à être adaptée à chaque application.

Les connaissances sémantiques d'une application sont généralement stockées dans une base de données. Une des approches du raisonnement des connaissances sémantiques est de les représenter sous forme de formules logiques. Ces formules contiennent des variables associées à des constantes et qui peuvent être saisies. Les logiques du premier ordre ou d'ordre supérieur peuvent être utilisées. Les concepts contenus dans le signal sont associés à un processus d'interprétation. De nouvelles assertions peuvent être obtenues par un processus d'inférence à partir de nouveaux concepts.

Les représentations dépendantes de la tâche sont sujettes à une portabilité moindre vers de nouveaux domaines ou applications. Récemment, deux représentations ont fait leur apparition : FrameNet et PropBank.

La sémantique informatique doit être basée sur des modèles qui représentent les connaissances avec des schémas incluant les hypothèses des entités sémantiques. Ces schémas appliquent des relations entre les signaux et la compréhension.

Dans ce but, il existe des classes d'objets, appelées *Frames* [Baker *et al.*, 1998]. Les Frames ont une structure identifiée par un nom et un jeu de paire rôle-valeur, appelés *slots*. Des procédures et fonctions sont attachées aux *slots*. Une Frame peut être vue comme une organisation de concepts. Beaucoup de représentations sémantiques sont basées sur des boîtes à n-relations entre les concepts, telles que le propose Filmore et comme utilisées dans le projet FrameNet [Baker *et al.*, 1998]. Les systèmes à représentation par boîte sont constitués de peu de boîtes, chacune d'entre elles représentant une contrainte sémantique élémentaire. Les Frames utilisées dans la sémantique sont inspirées des structures à base de boîtes [Fillmore, 1968] et ont été considérées comme des structures cognitives. Dans l'article [Jackendoff, 2002], la théorie exposée présente une structure sémantique obtenue par composition de fonctions.

FrameNet a pour but de documenter la réalisation syntaxique des arguments des prédicats du lexique général anglais en annotant des rôles sémantiques dans un corpus. Le projet est focalisé sur des Frames sémantiques, indépendantes de la tâche, qui sont définies comme représentation sémantique de la situation, elle-même issue des différentes associations de Frame. D'abord, les Frames et les Frame-éléments associés sont définis, ensuite des exemples de phrases, issues du British National Corpus, sont annotées. Considérons la phrase suivante : « Le client accepte le contrat». L'action d'intention du verbe accepter est représentée avec la Frame présentée ici : 2.4.

Nous pouvons voir un autre exemple d'annotation en Frame dans le cadre du projet européen LUNA (encadré 2.5). Cet exemple, issu du corpus `MEDIA`, est au format FrameNet.

Une ontologie spécifique définit cette Frame comme un prototype sur laquelle plusieurs instances peuvent être obtenues. Dans ces instances, les valeurs sont associées aux rôles. Entre crochets sont représentées les contraintes associées aux valeurs de

{accepte

Définition :	verbe
Agent :	[humain]
Thème :	[...]
...	
Autres rôles :	[...]}

TAB. 2.4: *Exemple de Frame*

```
<frame frname="LOCATION" id="t_1015_2_spk_Frame_1275" concept="location" span="marseille" semid="-1">
        <frameelement fename="location_town" id="t_1015_2_spk_FrameElement_848" frameid="t_1015_2_spk_Frame_1275" semid="-1 " concept="location_town" span="marseille" value="marseille" subframeid=""/>
</frame>
```

TAB. 2.5: *Exemple d'une Frame extraite du corpus* `MEDIA` *dans le cadre du projet LUNA*

chaque rôle. Les représentations spécifiques peuvent être attachées à des slots comme des termes généraux, modifications circonstancielles, informations probabilistes, degrés de certitudes, temps ou chronologie. Trouver les valeurs de chaque rôle peut être vu comme un processus de complétion des slots effectué par des procédures associées.

PropBank est un autre projet qui additionne des calques d'informations d'arguments prédits ou de labels sémantiques à la structure syntaxique de Penn Treebank. Les labels des rôles sémantiques indiquent le rôle de chaque argument, pour toutes les cibles prédites dans une phrase (les verbes, par exemple). Contrairement à FrameNet, PropBank nécessite une analyse syntaxique sur laquelle les rôles sémantiques vont être attribués. L'article [Tur *et al.*, 2005] présente une méthode pour faire du *bootstrap* sur les systèmes de compréhension de la parole basés sur PropBank. Les ensembles prédicats-arguments contribuent à former une armature quand la frame résultante a une signification spécifique. Pour certaines applications la seule composition utile est une frame représentant les actes de dialogue dont les composants sont des éléments sémantiques. Le but de ces applications peut être représenté par des frames qui contraignent l'agrégation des paires prédicats-arguments à des actions spécifiques du système.

2.3.5 D'une séquence de mots à une représentation sémantique

La représentation sémantique est basée sur l'application de relations entre les symboles et leur signification. Le processus peut être vu comme un processus de traduction des phrases en représentation sémantique. La question importante concerne la définition des symboles. Si on suppose que ces symboles sont des mots, alors il s'agit de traduire des symboles en langage naturel. Des systèmes de compréhension de la parole sont basés sur cette supposition. Dans ce cas, l'interprétation du message vocal est similaire à l'interprétation d'un message écrit. Un analyseur sémantique peut intervenir avec un analyseur syntaxique afin de produire une représentation sémantique d'assez bonne qualité pour un système à déduction logique. En effet, chaque élément

syntaxique important de la phrase correspond à un élément conceptuel, mais l'inverse est faux.

Les modèles à base de grammaires manuelles

La modélisation basée sur l'utilisation de grammaires manuelles est mise au point par des experts en linguistique. Les systèmes utilisant cette approche sont des systèmes qui ont besoin d'une ontologie particulièrement précise et qui doit tenir compte de tous les cas possibles pouvant être rencontrés lors de l'interprétation d'un message. Ce type de système ne requiert aucun corpus annoté pour la création des modèles. Ces systèmes utilisent généralement un segmenteur syntaxique sur lequel sera basé une projection sémantique.

Une analyse syntaxique peut être produite par un segmenteur qui produit un arbre de dérivation associant une phrase et des labels sémantiques, les prédicats et les arguments, qui sont associés à chaque élément de cet arbre. L'exemple : « Le client accepte le contrat » donne le résultat présenté dans la figure 2.2 avec un arbre de dérivation auquel sont attachés les labels sémantiques.

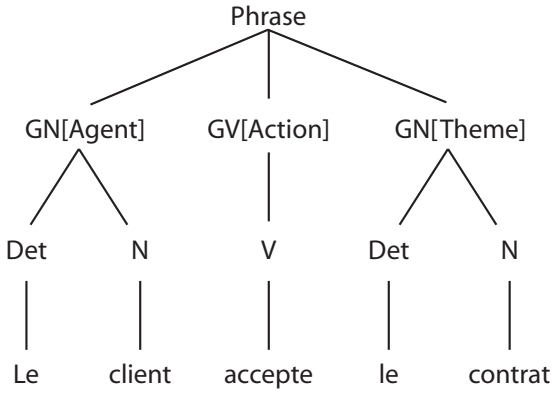

FIG. 2.2: *Arbre de décomposition associé à des labels sémantiques*

Il y a différentes approches pour associer les labels sémantiques et les symboles non-terminaux. La propriété des structures est utilisée pour imposer des contraintes sur les règles de grammaire. Ces grammaires suppriment les combinaisons impossibles ; par exemple, autoriser l'association entre un sujet groupe nominal (GN) et un groupe verbal (GV) d'une même phrase. Par l'exemple : « *Donner la liste de tous les employés des compagnies basés/basées au centre-ville* », les propriétés syntaxiques de *basés* et de *basée* produisent une analyse syntaxique différente :

– la liste des employés basés dans le centre-ville
– la liste des employés des compagnies qui sont basées au centre ville.

Une association de formules de construction sémantique avec une analyse syntaxique est proposée dans les grammaires de catégories. Ces dernières sont conçues pour obtenir une représentation sémantique de surface [De Mori, 1998].

La syntaxe de la langue peut parfois être assimilée à de l'algèbre et les catégories grammaticales à des fonctions. Les représentations lexicales sont alors associées à des symboles syntaxiques qui suggèrent la possibilité de continuer l'analyse syntaxique et de générer l'expression sémantique. Les connaissances sémantiques sont associées dans ce cas-là à des entrées lexicales et les formules logiques sont composées durant le découpage par déclenchement d'événements. Pour approfondir le sujet des règles de dérivation de l'interprétation sémantique à partir d'arbres syntaxiques, voir l'article [Allen, 1988].

Ces systèmes sont performants sur les textes bien structuré (i.e. romans, articles de journaux, etc.), mais ils perdent leur robustesse lorsqu'on les applique à la langue naturelle (parole, textes issus du web, langage familier, langage SMS...). En effet, la langue naturelle prend beaucoup de libertés structurelles et lexicales (répétitions, disfuences, etc.) comme vu dans la section 2.3.3. Ces variations dans la langue ne peuvent être prises en compte par les modèles formels qu'au prix d'une sur-génération de règles, car ces derniers ne peuvent considérer un évènement ou une séquence d'évènements que si et seulement s'ils ont été pris en compte dans les règles manuelles.

Comme pour la Reconnaissance Automatique de la Parole (section 2.2.2), représenter une langue naturelle de manière formelle reste une tâche difficile et longue. La modélisation stochastique de la langue est privilégiée dans le cadre d'applications de dialogue oral, en particulier lorsque celles-ci permettent à l'utilisateur de s'exprimer librement. Dans le cadre d'applications complexes utilisant un processus de comprehension avec une ontologie riche de plus d'une centaine de concepts comportant de l'ambiguité, le systèmes à base de corpus ont démontré leur efficacité lorsque ces systèmes sont utilisés pour la compréhension de la parole notamment pour des application comme ATIS (Voir section suivante). Cependant dans le cadre d'applications simples, dont le domaine sémantique est très restreint (moins d'une cinquantaine de concepts avec pas ou peu d'ambiguité), les modèles utilisés sont généralement à base de règles manuelles. Enfin, les problématiques liées au coût de collecte et de création de corpus d'apprentissage des systèmes à base de corpus font que les systèmes à base de règles manuelles sont généralement utilisés dans les systèmes actuellement déployés, même si ces derniers sont complèxes.

Des modèles de compréhension dépendant de la tâche

Beaucoup de problèmes liés à la l'interprétation automatique dans les systèmes de compréhension sont issus du fait que beaucoup de phrases sont agrammaticales, la partie *RAP* fait des erreurs dans les hypothèses de mots, ce qui entraîne une limitation de la couverture des grammaires. Ces considérations suggèrent que plus le système est spécifique, plus les modèles sont robustes aux énoncés sémantiques.

Au début des années 90, la DARPA (Defense Advance Research Projects Agency) a fondé ATIS (Airline Travel Information System). Le projet a permis l'apparition d'un certain nombre de systèmes de compréhension, dépendants de la tâche. Les données collectées avec un système de requêtes données par l'utilisateur à propos des informa-

tions et réservations de transport aérien ; par exemple : *je voudrais aller de Boston à New York la semaine prochaine* ou *est-ce que le repas est prévu avec ce vol ?*

ATIS fournit un test de référence pour de nombreux systèmes de compréhension, qu'ils soient à base de grammaires manuelles, statistiques ou des systèmes hybrides. L'élément de Frame contient typiquement les informations de départ, d'arrivée, les villes et les dates associées.

L'analyseur linguistique TINA [Seneff, 1989] a été développé au MIT. 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. Les probabilités liées aux règles sont estimées sur un corpus d'apprentissage. Elles servent à contraindre la recherche pendant l'analyse car, sans elles, toutes les segmentations sont possibles. Afin d'améliorer la robustesse du système, des analyses partielles sont possibles. Dans ce cas, l'analyseur produit l'ensemble des analyses de chaque mot, chaque analyse pouvant commencer à ce même mot du message.

Les modèles statistiques de compréhension

Outre le découpage partiel et le *back-off* sur certaines correspondances, Lorsque le découpage partiel échoue, il est utile de d'avoir des outils pour représenter des connaissances caractérisées par les probabilités distribuées [De Mori, 1998]. On peut obtenir automatiquement les paramètres de ce modèle en les apprenant à partir de corpus annotés manuellement. Ce travail est aussi long que complexe, fiable et dispose d'une bonne couverture.

Le système Chronus d'`AT&T` [Pieraccini *et al.*, 1991] est basé sur le paradigme du canal bruité, communément utilisé pour formaliser le problème de la reconnaissance vocale. Avec ce système, les connaissances sémantiques sont représentées par des modèles de Markov, dans lesquels les observations sont les mots w, utilisant un état caché pour chaque concept sémantique.

Une représentation sémantique sous forme d'arbre est proposée par le Hidden Understanding Model (HUM) [Miller *et al.*, 1994]. Un exemple de cette représentation est présenté en figure 2.3. Les éléments sémantiques correspondant à une décomposition partielle en arbre sont utilisés pour identifier les symboles non-terminaux de la grammaire conceptuelle hors-contexte (SCFG). Le modèle de langage conceptuel est représenté sous forme d'arbre : les concepts sont les nœuds, avec les sous-concepts comme des nœuds-fils. Chaque nœud terminal est le parent d'un mot ou d'une séquence de mots. Il est important de noter que l'analyse d'une phrase entière n'est possible que si les limites de celle-ci sont connues.

Les systèmes Chronus et HUM de BNN sont présentés en détail dans [Wang *et al.*, 2005] dans lequel est aussi décrit un système hybride à base de modèle de Markov et de SCFG. Dans cette approche, une grammaire hors-contexte est manuellement écrite. Ces grammaires concernent tout ce qui n'est pas spécifique au domaine étudié, tel que

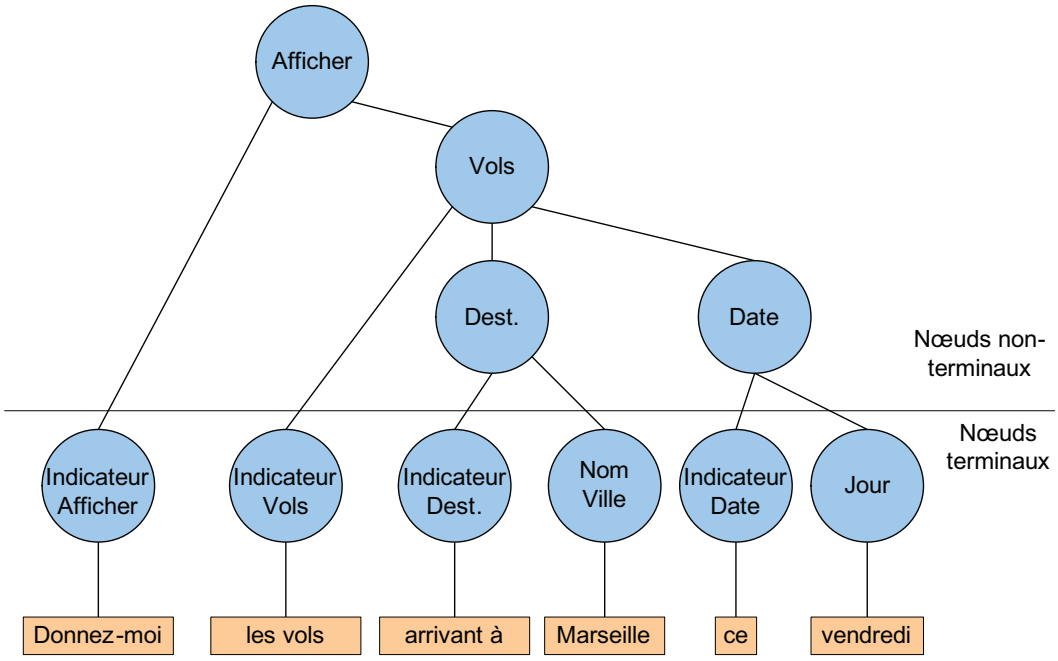

FIG. 2.3: *Un exemple de représentation sémantique hierarchique issu de ATIS*

les noms de ville, les dates, les nombres, etc. Ces règles de grammaires ainsi apprises peuvent être enrichies par des bases de données ou pré-remplies par des librairies.

Dans le système Chanel, les premières entités générales comme les noms de villes d'aéroport et les dates sont marqués. Leur rôle est déterminé par un arbre de classification automatique sémantique (SCT) appris automatiquement sur des données d'entraînement. Un exemple simplifié est présenté figure 2.4 déterminant le rôle d'une ville. Les instances de Frame sont composées par des règles issues de ces hypothèses de rôle (*cf.* chapitre 14 de [De Mori, 1998]).

À l'université de Cambridge [He et Young, 2006], proposent une approche basée sur les SCFG qui ne nécessite des données que partiellement annotées. La solution utilise des modèles de vecteurs à états cachés (HVS). Chaque état insère un arbre contextuel dans un vecteur (ou une pile). Ensuite, l'état de la transition de chaque mot émis est factorisé en n opérations d'extraction et une opération d'insertion de pile pour ce même mot. Cependant, le modèle peut seulement modéliser les grammaires hors-contexte du coté droit. Les probabilités des transitions pour chaque n sont apprises sur le corpus ou sous forme de modèles grammaticaux. La profondeur de la pile et le nombre n d'opérations d'extraction sont limités pour des raisons d'efficacité.

L'utilisation des grammaires sémantiques est abordée dans l'article [Wang *et al.*, 2005]. Plus récemment, des modèles statistiques combinant syntaxe et sémantique ont été proposés. La figure 2.5 présente un arbre syntaxico-sémantique de la phrase : « Donner la liste des vols de la TWA de Washington à Philadelphie ». Pendant que la segmentation

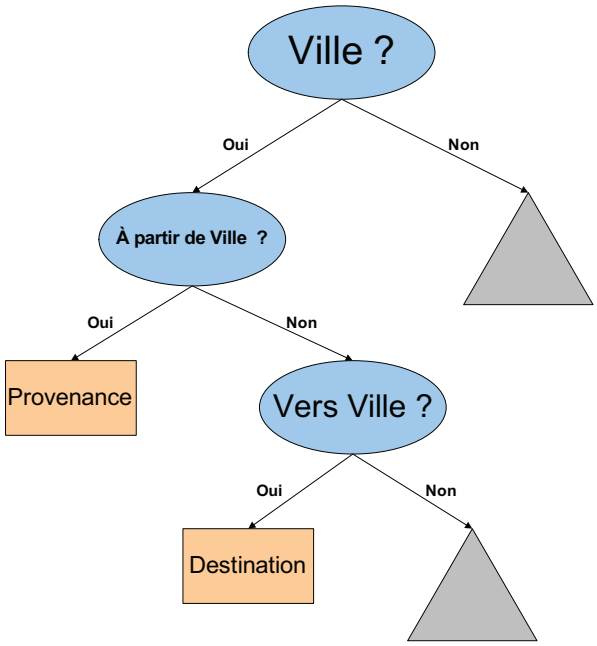

FIG. 2.4: *Arbre de classification automatique sémantique (SCT) pour un domaine d'ATIS*

est extraite depuis la phrase, grâce à un segmenteur anglais général, les sous-arbres sont annotés semi-automatiquement avec les caractéristiques propres de l'attribut *VilleOrigine* (avec les lignes entières) et contextuel (en pointillés) [Moschitti *et al.*, 2007]. Avec cette approche, les hypothèses sémantiques sont attachées dynamiquement à des symboles non-terminaux, issus d'une grammaire syntaxique générale. Elle est préférable à des grammaires sémantiques statiques ayant la possibilité de réécrire les symboles sémantiques non-terminaux.

Les modèles de classification pour la compréhension

La robustesse de la compréhension de la parole spontanée a largement été atteinte dans la classification d'énoncés de dialogue par rapport à un contexte ou dans la caractérisation des intentions d'un utilisateur. Ces types de travaux ont été effectués par AT&T sur *How May I Help You ?* [Gorin *et al.*, 1997]. Les conversations des utilisateurs ont été classées en un nombre prédéfini d'intentions. L'approche choisie utilise des classifieurs discriminants (algorithmes de *Boosting* ou *SVM*) avec des modèles *n*-grammes de mots et des entités nommées spécifiques à la tâche, apparaissant dans l'énoncé comme des caractéristiques [Gupta *et al.*, 2006]. La combinaison de connaissances écrites manuellement et des résultats de l'apprentissage automatique des connaissances sémantiques est proposée dans [Shapire *et al.*, 2005]. Des alternatives à l'utilisation de classifieurs à base de SCT, d'algorithmes de *boosting* et de *SVM*, sont proposés dans [Raymond *et al.*, 2007] afin d'améliorer la robustesse de la classification.

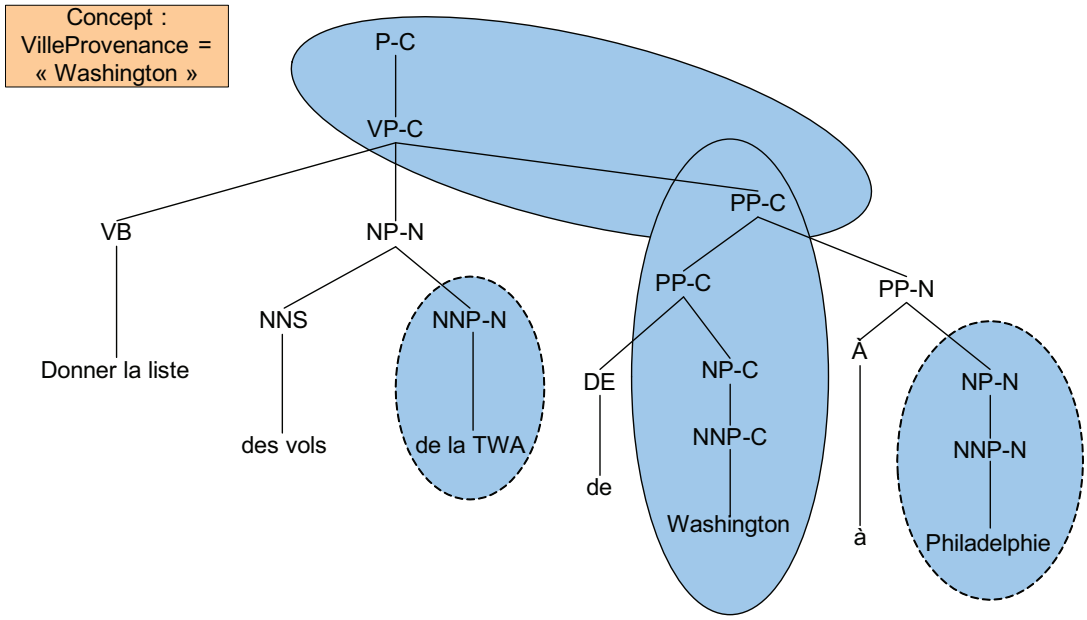

FIG. 2.5: *Exemple d'un arbre syntaxico-sémantique du domaine d'ATIS*

Pour d'autres systèmes discriminants [Papieni *et al.*, 1998], les modèles statistiques de traduction sont utilisés pour traduire une phrase source S en un MRL cible T en maximisant la probabilité $P(T|S)$. La tâche principale de l'apprentissage est de déterminer les corrélations entre des séquences de mots dans un langage et les entités de l'ontologie. Lorsque le canal source échoue à capturer une telle corrélation, alors un modèle est construit pour récupérer les probabilités *a posteriori* $P(T|S)$. L'utilisation de modèles discriminants pour cette approche est décrite dans l'article [Raymond et Riccardi, 2007].

2.3.6 Collecte de corpus et comparaison de différentes approches.

Sans tenir compte de la méthode choisie pour effectuer le processus de traduction d'un signal vocal en une interprétation sémantique, des corpus de développement et d'évaluation sont nécessaires pour le développement et l'évaluation des systèmes. Pour les systèmes à base de règles manuelles, les corpus sont utiles pour écrire les règles, grâce à des exemples de parole et pour vérifier leur couverture. S'agissant des systèmes aux méthodes basées sur des corpus, les corpus annotés sémantiquement sont utilisés pour l'apprentissage. Concernant l'évaluation des systèmes, les deux approches ont besoin de corpus de référence annotés manuellement.

La disponibilité de corpus de parole pour la compréhension est l'une des difficultés les plus importantes pour créer des modèles sémantiques. En effet, les coûts de collecte et d'annotation manuelle sémantique d'un corpus de parole sont particulièrement importants.

Les corpus liés à des applications spécifiques ne sont généralement pas du domaine public. Cependant, deux exemples notables de corpus de dialogues vocaux annotés sémantiquement et disponibles publiquement sont le corpus ATIS (mentionné précédemment) et le corpus MEDIA, un corpus de dialogue téléphonique sur des réservations et informations touristiques collectées par la méthode du Magicien d'Oz. Les descriptions, discussion et expériences sont décrites dans le chapitre 14 de [De Mori, 1998], pour le premier et, pour le second, dans la section 2.4.1 et dans [Raymond et Riccardi, 2007].

Il n'y a pas d'approche clairement supérieure aux autres dans ces évaluations. Les méthodes statistiques ont prouvé leur robustesse au prix d'un large corpus d'apprentissage annoté, ce qui est rarement possible. Parmi les méthodes statiques proposées pour la compréhension, les méthodes discriminatives, telles que les Conditionnals Randoms Fields (CRF), fournissent un avantage sur les méthodes génératives comme reporté dans [Raymond et Riccardi, 2007] sur le corpus MEDIA. Cependant, comme souligné dans ce papier, les méthodes discriminantes semblent être plus sensibles aux erreurs et au manque de consistance dans les corpus d'apprentissage, que les méthodes génératives.

2.3.7 Traiter les hypothèses multiples du module de reconnaissance de la parole

Il convient de noter que l'approche la plus intuitive est une approche dite séquentielle. Cette approche du problème de la compréhension de la parole se base sur une extraction de la meilleure hypothèse du module de reconnaissance de la parole qui est ensuite transmise au système (ou module) de compréhension de la parole. Mais cette approche, bien que plus aisée à mettre en œuvre, présente l'inconvénient de ne conserver qu'une hypothèse, laquelle peut contenir des erreurs de reconnaissance. Ces erreurs sont répercutées ensuite sur le module de compréhension. Ce dernier ne pourra pas les traiter, faute de connaître les autres hypothèses.

Un des moyens de traiter les erreurs de reconnaissance est de tenir compte d'un jeu d'hypothèses, au lieu de ne sélectionner que la meilleure hypothèse. Cet ensemble se retrouve sous la forme d'un graphe de mots ou de liste des n-meilleures hypothèses. Lorsque les hypothèses du graphe de mots sont produites, il est probable que l'hypothèse attendue soit présente dans le graphe, ce qui rend possible une hypothèse sémantique cohérente avec les hypothèses du module de la reconnaissance de la parole. De nombreux algorithmes ont été proposés pour générer ces graphes d'hypothèses probabilistes sémantiques à partir d'un graphe d'hypothèses probabiliste de mots ou lors de la reconnaissance de la parole.

AT&T propose une approche intégrée : un graphe de mots est soumis en entrée du module de décodage conceptuel [Bangalore et Johnston, 2004]. Il y est décrit un mélange de modèles de langage pour des applications multimodales, avec une partie entraînée sur des données propres au domaine et d'autres obtenues par génération à partir de grammaires. Comme vu dans la partie 2.3.5, le processus de compréhension est la reconnaissance d'une séquence de prédicats ou de concepts qui maximisent la probabilité

$P(T|W)$, où T est la séquence de concepts et W est la séquence de mots. L'approximation est faite grâce à des *bi*-grammes ou des *tri*-grammes.

Chez IBM [Sarikaya *et al.*, 2005], le système proposé génère une liste des n-meilleures hypothèses de mots, lié à un modèle de langage (LM) *tri*-gramme dépendant de l'état du dialogue. Cette liste est réévaluée avec deux modèles sémantiques. Une grammaire intégrée (EG) hors-contexte sémantique est définie pour chaque concept. Chaque grammaire fait ressortir le concept associé et recherché dans la phrase. Les probabilités *tri*-grammes sont utilisées pour évaluer les différentes hypothèses produites par le modèle de grammaire. Un autre modèle de langage, appelé *maximum entropy* LM (MELM), calcule les probabilités d'un mot à partir de l'historique donné par le modèle de *maximum entropy*.

Nous décrirons dans le chapitre 3 notre système qui utilise une approche intégrée du décodage conceptuel. Puis, nous comparerons les approches intégrées et séquentielles dans les expériences des chapitres 4 et 5.

2.3.8 Mesures de confiance sémantique

Les systèmes de reconnaissance de la parole et de compréhension de la parole, dits « état de l'art », font des erreurs qui doivent être identifiées, afin d'appliquer les stratégies appropriées pour l'exécution d'actions, telles que la correction d'erreurs dans les dialogues homme-machine. Les probabilités *a posteriori* $P(I|Y)$, d'une interprétation I donnée pour la séquence de paramètres acoustiques Y, ne constituent pas le meilleur indice de fiabilité des hypothèses. L'indice de fiabilité doit tenir compte des informations qui ne sont pas codées dans Y, comme la cohérence des différentes hypothèses disponibles avec l'historique du dialogue, incluant les énoncés systèmes et des corrections.

L'estimation de la fiabilité d'une interprétation relève de certains points : choisir la couverture de la mesure de confiance (mots, concepts ou énoncés), définir le jeu de paramètres pris en compte dans l'évaluation (les paramètres de la reconnaissance, ceux de la compréhension, le contexte de dialogue), la combinaison efficace des différents paramètres et le choix d'une stratégie de décision qui tienne compte de tous les paramètres obtenus.

La fiabilité des modèles de langage n'étant pas l'un des sujets d'étude, nous n'aborderons pas ce sujet dans cette thèse. Néanmoins, le domaine de la fiabilité des modèles de langage est une des ouvertures possibles à la suite du travail présenté ici.

2.4 La campagne d'évaluation Technolangue MEDIA/EVALDA

2.4.1 Le corpus MEDIA

Les applications de dialogue Homme-Machine considérées dans ce chapitre 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 de base 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 à des réservations d'hôtel. Un corpus de 1 250 dialogues a été enregistré par ELDA selon un protocole de *Magicien d'Oz* (*WOZ*) : 250 locuteurs ont effectué chacun 5 scénarios de réservation d'hôtel avec un système de dialogue simulé par un opérateur humain. 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]. Un exemple de dialogue est donné entre un locuteur (**spk**) et compère (**woz**) dans l'encadré 2.6.

woz	Bienvenue sur le serveur MEDIA système d' informations touristiques et de réservation d' hôtel. Quelle information désirez-vous ?
spk	Je voudrais réserver pour le troisième week-end de janvier.
woz	Dans quelle ville souhaitez-vous réserver un hôtel pour le troisième week-end de janvier ?
spk	Dans un hôtel à Evreux dans le centre-ville.
woz	Quel type de chambres désirez-vous réserver ?
spk	Alors euh il me faudrait deux chambres doubles et une simple.

TAB. 2.6: *Exemple de dialogue extrait du corpus MEDIA*

2.4.2 Ontologie MEDIA

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

- tout d'abord, une paire attribut-valeur, correspondant à une représentation sémantique *à plat* d'un énoncé ;
- puis, un spécifieur permettant de définir des relations entre les attributs et qui 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*
16	pour un couple	sejour-nbCouple	+		1
17	et un enfant	sejour-nbEnfant	+		1
18	comptez au niveau	null	+		
19	des prix	objet	+	chambre	paiement-montant-entier
20	soixante	paiement-montant-entier	+	chambre	60
21	euros	paiement-monnaie	+		euro
22	par chambre	null	+		

TAB. 2.7: *Exemple de message annoté hors-contexte du corpus MEDIA*

Le tableau 2.7 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. Par exemple, *nom-hotel* dérive de la table *hotel* contenue dans la base de données et du champ associé au nom de l'hôtel.

Les spécifieurs sont dépendants de l'attribut et permettent de modifier ou de préciser la signification associée à l'attribut suivant le cas. Dans exemple 2.7, le spécifieur *chambre* associé à l'attribut *paiement-montant-entier* permet de préciser que la chaîne W^{c_n} annotée *paiement-montant-entier-chambre* se rapporte au montant de la réservation en cours : *soixante*.

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. Dans ce chapitre, les 4 premiers lots composent le corpus d'apprentissage (soit 720 dialogues contenant environ 12 000 messages d'utilisateurs) et les lots 5 et 6 forment respectivement les corpus de développement et de test contenant 200 dialogues avec 3 000 messages d'utilisateurs chacun.

2.4.3 Annotation

La première partie de la campagne d'évaluation MEDIA consistait à analyser des tours de dialogues sans tenir compte du contexte du dialogue. De ce fait, les références ne sont pas résolues. L'exemple proposé dans le tableau 2.8 illustre parfaitement le cadre *hors-contexte*, en donnant aux liens référentiels annotés *lienRef* les valeurs *singulier* ou *pluriel*. Ce type d'annotation vise principalement l'attribut *lienRef* et n'influence pas la résolution des autres attributs c_n pour chaque W^{c_n}.

Par opposition au *hors-contexte*, la résolution *en-contexte* tient compte du contexte du dialogue. Ceci a pour conséquence la résolution des références, qui passe d'une annotation simplifiée (les valeurs *singulier* ou *pluriel*) à une annotation plus complète, avec pour valeur, à présent, les identifiants n des W^{c_n} associés. De plus, nous pouvons avoir plusieurs identifiants pour une seule référence. Les références sont associées aux attributs *lienRef* et les spécifieurs associés sont :

- coRef : désigne le référent par référence directe ou coréférence et regroupe les pronoms définis et démonstratifs ;
- coDom : désigne le référent grâce à une marque linguistique d'altérité ;
- elsEns : désigne le référent d'après des propriétés sémantiques qui l'opposent à d'autres entités dans un ensemble ; on retrouve, dans ce cas là, les ordinaux, les superlatifs, les expressions relatives et certains pronoms démonstratifs.

Voici un exemple d'annotation *hors-contexte* annoté *en-contexte* dans le tableau 2.8

Annotation *hors-contexte*						
n	W^{c_n}	c_n		*mode*	*spécifieur*	*valeur*
0	est	null		+		null
1	il	lienRef		+	coRef	singulier
2	proche du	loc-distanceRelative		?		proche
3	centre-ville	loc-lieuRelatif		?	general	centreville
Annotation *en-contexte*						
n	W^{c_n}	c_n	référence	*mode*	*spécifieur*	*valeur*
0	est	null		+		null
1	il	lienRef	palais	+	coRef	singulier
2	proche du	loc-distanceRelative		?		proche
3	centre-ville	loc-lieuRelatif		?	general	centreville

TAB. 2.8: *Exemple de message annoté hors-contexte annoté en-contexte du corpus* `MEDIA` *sur le message « Est-il proche du centre-ville ?» énoncé après le prompt : « Je vous propose l'hôtel du Palais»*

Enfin, le passage entre les deux cadres est aussi marqué par un ajout de spécifieurs, relatifs aux événements apparus précédemment dans le dialogue. Ainsi, au tableau 2.7 annoté *hors-contexte*, il nous faut ajouter les spécifieurs comme présenté dans le tableau 2.9 :

- *reservation* aux concepts *sejour-nbCouple* et *sejour-nbEnfant* ;
- conserver *chambre* aux concepts *objet* et *paiement-montant-entier*.

On retrouve des spécifieurs faisant référence à des contextes énoncés dans les tours précédents.

Le travail réalisé de lors de la campagne `MEDIA` s'est limité à la représentation *hors-contexte*. Ainsi tous les travaux présentés dans ce chapitre sont basés sur l'évaluation *hors-contexte*.

n	W^{c_n}	c_n	*mode*	*spécifieur*	*valeur*
16	pour un couple	sejour-nbCouple	+	reservation	1
17	et un enfant	sejour-nbEnfant	+	reservation	1
18	comptez au niveau	null	+		
19	des prix	objet	+	chambre	paiement-montant -entier
20	soixante	paiement-montant-entier	+	chambre	60
21	euros	paiement-monnaie	+		euro
22	par chambre	null	+		

TAB. 2.9: *Reprise de l'exemple de message annoté hors-contexte 2.7, cette fois annoté en-contexte.*

2.4.4 Accord inter-annotateurs

Afin de vérifier la qualité du corpus, une évaluation sur l'accord entre les différents annotateurs a été faite. Cet accord est appelé *l'accord inter-annotateur (ou IAG)*. La mesure utilisée permettant de valider est la mesure *kappa* telle que [Cohen, 1960] :

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.19}$$

avec $P(A)$ correspondant au nombre de fois où les annotateurs sont d'accord par rapport au nombre d'annotations 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.

IAG	1	2	3	4	5	6	7
Nbr Dialogues	20	8	10	10	10	10	10
Nbr Tours speaker	310	125	183	165	137	106	163
Nbr Segments	667	459	478	372	455	342	459
Accord (%)	66.1	76.2	78.4	**89.5**	**83.1**	**83.9**	**87.8**

TAB. 2.10: *Chiffres de l'accord inter-annotateur (IAG) présentant l'évolution de celle-ci en fonction des différents lots, apprentissage (1, 2, 3 et 4) développement (5 et 6) et de test (7)*

Lors de la mise au point du corpus MEDIA, les mesures d'IAG effectuées sont d'environ 90% [Bonneau-Maynard *et al.*, 2005]. Cette évaluation tend à montrer que les annotations sont assez homogènes et donc à valider le corpus pour la campagne d'évaluation MEDIA.

Nous pouvons voir dans le tableau 2.10 les résultats des différents IAG. Ces résultats sont satisfaisants à partir de la quatrième mesure d'IAG, celle-ci correspondant au moment où les annotatrices ont eu fini d'annoter les 3 premiers corpus d'apprentissage. Les dernières IAG ont été faites sur le corpus de développement et de test, afin de vali-

der les résultats obtenus par les participants à la campagne d'évaluation. Il est à noter qu'à chaque modification de l'ontologie, les corpus déjà annotés ont été révisés suivant les modifications apportées.

2.4.5 Corpus audio

Le corpus audio MEDIA a été enregistré avec un système de Magicien d'Oz (*WOZ*) ; un système WOZ consiste à simuler un serveur vocal par l'intermédiaire d'une personne. Les utilisateurs qui appelaient le service ont cru appeler un serveur de renseignement touristique. Il s'agissait en fait d'humains qui simulaient les réactions d'une machine en fonction d'un scénario.

Les 1 257 dialogues ont été enregistrés avec 250 locuteurs différents qui avaient 5 scénarios différents à jouer à chaque fois. Les contraintes acoustiques étant minimes, le corpus a été enregistré avec toutes sortes d'appareils téléphoniques (téléphones fixes, téléphones cellulaires, etc.), ce qui donne une grande variété de qualités acoustiques, qui est très proche de la réalité du système et de sa mise en œuvre.

Dans les expériences décrites plus loin, nous n'utiliserons que le corpus de test *hors-contexte* pour mesurer l'efficacité de notre système, conformément aux expérimentations menées lors de la campagne MEDIA.

2.4.6 Systèmes développés pour la campagne MEDIA

Les différents systèmes qui ont participé à la campagne sont basés sur différentes approches. Celui du LIMSI (équipe 1) est un système stochastique basé sur les corpus d'apprentissage, celui du LORIA et du VALORIA utilisent des méthodes basées sur des règles manuelles, enfin celui du LIMSI (équipe 2) est un système mélangeant les deux méthodes précédentes.

Systèmes à base de connaisances

Le système du LORIA
L'approche utilisée dans le système du LORIA[2] est basée sur la décomposition et la description logique :

- un parseur LTAG [Crabbe *et al.*, 2003] produit une analyse syntaxique ;
- un constructeur sémantique compositionnel produit un graphe conceptuel à partir de l'analyse syntaxique ;
- un module de projection aplatit le graphe et construit le format sémantique voulu, c'est à dire une paire d'attribut-concept dans le formalisme MEDIA et ordonné selon la phrase à décoder.

[2]Pour plus d'informations, voir [Denis *et al.*, 2006].

Ce système n'effectue aucun apprentissage et n'a besoin d'aucun corpus annoté, mais requiert un guide d'annotation de qualité. De plus, il a besoin de :

- un lexique morphologique extrait de *Multext Lexicon* (5 400 mots et 3 000 lemmes) ;
- un lexique syntaxique créé pour les heuristiques simples (*e.g.* : les noms sont liés à l'arbre des noms) ;
- une grammaire au format LTAG (80 arbres) ;
- un lexique sémantique utilisé pour produire les graphes conceptuels (150 schémas) ;
- une ontologie interne pour filtrer le graphe conceptuel ;
- une ontologie externe dont les concepts sont définis en termes de concepts internes (130 concepts).

Le système du VALORIA

Le Système LOGUS utilise une approche logique de la compréhension de la langue naturelle en français [Villaneau *et al.*, 2004]. Ce système, à base de grammaires manuelles, utilise des formules logiques semblables à la logique de D. Vanderveken. L'analyse syntaxique produite est utilisée comme base de l'analyse conceptuelle. Les concepts et les structures conceptuelles sont utilisés pour créer la formule logique telle qu'elle soit transformable en un graphe conceptuel. Les deux analyses précédentes sont faites séquentiellement. La principale difficulté de ce type d'analyse est la mise au format MEDIA du résultat de l'analyse sémantique.

Systèmes à base de corpus

LIMSI équipe 1

Le système du LIMSI-1 utilise une méthode stochastique à base de corpus. Le modèle sémantique stochastique initial à deux niveaux a été étendu à un niveau 2+1, où un niveau stochastique supplémentaire est chargé de normaliser les valeurs au format MEDIA [Bonneau-Maynard et Lefevre, 2005]. Afin de produire le résultat final, le système comprend deux parties : d'une part, le décodage conceptuel, qui produit les séquences de mode-attribut associées aux séquences de mots, d'autre part, une traduction des séquences de mots en valeurs normalisées. Une variante a été développée où le décodage conceptuel et la phase de normalisation ont été découplés (d'où le niveau 2+1). Ce nouveau modèle a été complété avec 3 nouvelles techniques permettant une amélioration relative de 20% sur le corpus de développement : normalisation stochastique basée sur les pénalités, propagation du mode et recomposition hiérarchique.

Systèmes mixtes

LIMSI équipe 2

Le système du LIMSI-2 utilise la même approche que celle utilisée pour la détection d'actes de dialogues [Rosset et Tribout, 2005]. Il utilise des règles manuelles pour généraliser certains concepts (noms de villes, dates…) mais il requiert un corpus d'ap-

prentissage pour les modèles de langage et le système de classification. Les modèles de langages serviront à la segmentation de la phrase en utilisant des marqueurs symbolisant des frontières et le système de classification permettra le décodage conceptuel des morceaux ainsi segmentés. Le tout devra être normalisé au format MEDIA.

LIA
Il est à noter que notre système utilise conjointement une base de connaissance pour certains concepts génériques (dates, nombres...) et un corpus d'apprentissage pour les modèles conceptuels associés aux domaines spécifiques de l'application. Ceci nous place dans la catégorie des systèmes mixtes. L'originalité du système est d'avoir été conçu pour prendre en entrée des graphes de mots issus du module de RAP et non du texte formaté. Nous utilisons un étiqueteur sémantique stochastique couplé à des grammaires de concepts que nous appliquons sur le graphe de mots. Enfin nous appliquons des grammaires de valeurs afin d'extraire les valeurs à partir des mots correspondant aux concepts. Enfin la sortie est normalisée au format MEDIA. Pour plus de détails, voir la description du système chapitre 3.

2.4.7 Une mesure d'évaluation : le taux d'erreur conceptuel ou *Concept Error Rate (CER)*

Les performances sont mesurées par rapport au taux d'erreur sur les paires attribut/valeur. Un concept détecté est considéré comme correct uniquement si l'attribut du concept et sa valeur normalisée sont corrects d'après la référence. En alignant la chaîne de concepts détectés automatiquement et celle présente dans la référence établie manuellement, on calcule le nombre de concepts corrects C, le nombre de concepts insérés I, le nombre de concepts omis O, ainsi que le nombre de concepts substitués S (soit au niveau de l'attribut, soit au niveau de la valeur). Si R est le nombre de concepts de la chaîne référence, le taux d'erreur, appelé le *Concept Error Rate* (CER), se calcule selon la formule :

$$CER = \frac{I + O + S}{R} \times 100 \tag{2.20}$$

2.4.8 Les outils d'évaluation

MEDIAVAL

Ce programme est un outil d'évaluation développé spécialement pour la campagne TECHNOLANGUE/EVALDA/MEDIA [Bonneau-Maynard *et al.*, 2006]. Le logiciel compare une référence et une hypothèse annotées au format MEDIA[3]. Il y a plusieurs types d'évaluations possibles qui sont combinatoires pour cette campagne :

– *Mode(4)* qui prend en compte tous les modes ;

[3] voir ontologie à la section 2.4.2

– *Mode(2)* ne prend en compte que les modes *positif* et *négatif*, sachant que les modes *interrogatif* et *si possible* (ces deux derniers étant mis de cotés) sont assimilés au mode *positif* ;
– *Relax* qui ne prend pas en compte les *spécifieurs* ;
– *Value* permet d'évaluer les valeurs des *attributs* ;
– *Full* qui évalue les attributs, les spécifieurs, les 4 modes et les valeurs.

MEDIAVAL calcule le *Concept Error Rate* (CER) tel que dans l'équation 2.20 en fonction des types d'évaluation effectuée. Les résultats de la campagne présentés dans la section 2.4.9 sont issus de MEDIAVAL.

SCLite

Ce programme est un outil d'évaluation de sorties des systèmes de *RAP*. Sclite fait partie du toolkit NIST SCTK. Le logiciel compare l'hypothèse (HYP) qui est la sortie de la *RAP* et la référence (REF) textuelle. Après comparaison entre la REF et l'HYP (processus appelé alignement), des statistiques sont produites et une collection de comptes rendus peuvent être produits pour résumer les performances de la *RAP*.

Le processus d'alignement consiste en deux parties :
– faire correspondre les REF et HYP ;
– effectuer un alignement de la référence et de l'hypothèse.

SCLite permet, entre autres, de calculer le taux d'erreur mots ou *Word Error Rate* (WER) et le *Concept Error Rate* (CER), comme dans l'équation 2.20.

2.4.9 La campagne d'évaluation

L'évaluation proprement dite s'est faite en deux phases : d'une part, un test sur le corpus de développement (environ 1 000 tours de paroles) qui a servi de répétition pour l'évaluation en avril 2005 ; d'autre part, l'évaluation finale effectuée sur le corpus de test (environ 3 000 tours de paroles) en juin 2005. Les tours ont été choisis aléatoirement et mélangés par ELDA, le coordinateur de la campagne. Il est utile de préciser que les évaluations ont été faites sur des transcriptions, afin que les participants ne soient pas pénalisés par les performances d'un système de reconnaissance de la parole.

Protocole d'évaluation de la campagne

Chaque participant disposait du corpus d'apprentissage, contenant environ 11 000 tours de parole entièrement annotés, du dictionnaire sémantique et d'un manuel d'annotations. Il y a en moyenne 4,8 mots par tour de parole dans ce corpus. Un lexique de 3 786 mots du corpus MEDIA a été fourni, ainsi que la liste des 667 valeurs ouvertes qui apparaissent dans le corpus. La moyenne des attributs par tour observée dans le corpus est de 2,7. 144 différents attributs ont été observés dans le corpus d'apprentissage.

Nous avons pu observer que l'annotation sémantique n'est pas figée et elle peut donner lieu à des variantes à travers l'annotation des corpus (paragrahe 2.4.4). Certaines variantes sémantiques peuvent être autorisées pour un même segment. Après les résultats, les participants étaient invités à proposer des modifications ou des alternatives à la référence. Cette partie, appelée « adjudications », a abouti à un vote de la part de tous les participants, afin d'obtenir un consensus sur ces propositions alternatives. Seules 179 demandes ont abouti à des annotations alternatives ; cela correspond à environ 6% du corpus de test, avec généralement 2 alternatives par segments.

Pour l'évaluation, MEDIAVAL a été utilisé (*cf.* section 2.4.8) en comparant les insertions, les substitutions, les délétions et le CER (*cf.* équation 2.20). Pour chaque segment nous avons évalué le triplet mode, attribut et valeur. Différentes évaluations ont été faites :

- *Full* où nous évaluions le triplet au complet ;
- *Relax* où nous ne considérions plus les *spécifeurs* ;
- *mode(4)* où tous les modes étaient pris en compte (*négatif*, *positif*, *interrogatif* et *si possible*) ;
- *mode(2)* où seuls les modes *négatif* et *positif* étaient pris en compte (sachant que les modes *interrogatif* et *si possible* étaient assimilés au mode *positif*).

Résultats

	Full		Relax	
	4 modes	2 modes	4 modes	2 modes
LIA	41.3	36.4	29.8	24.1
LIMSI-1	**29.0**	23.8	**27.0**	21.6
LIMSI-2	30.3	**23.2**	27.2	**19.6**
LORIA	36.3	28.9	32.3	24.6
VALORIA	37.8	30.6	35.1	27.6

TAB. 2.11: *Tableau présentant les résultats de la campagne d'évaluation MEDIA des différents participants. Les chiffres sont des taux d'erreur conceptuels (CER)*

Le tableau 2.11 montre les résultats des systèmes utilisés lors de la campagne hors-contexte, en fonction des différents modes d'évaluation. Il apparaît que les meilleures performances sont obtenues par un système mixte : le LIMSI-2. Les systèmes à base de corpus sont juste derrières. En revanche, les systèmes basés sur des règles manuelles se heurtent à des problèmes de projection ou de traduction dans le format MEDIA. Les résultats ont été obtenu à partir du décodage des transcriptions manuelles du corpus MEDIA.

Enfin, la contre-performance de notre système en mode *full* lors de la campagne officielle s'explique par l'absence de traitement sur les spécifieurs. On remarque, en effet, qu'en mode *relax* lorsque l'évaluation ne tient pas compte des spéfieurs, les performances sont dans la moyenne des autres systèmes présents lors de la campagne. Le système a été adapté à la tâche par la suite, nous pouvons retrouver les résultats de notre système modifié dans la partie 4.1.

Chapitre 3

Décodage conceptuel et apprentissage automatique sur corpus annoté

Sommaire

3.1 Introduction

Dans les applications de dialogue Homme-Machine téléphonique, le processus d'interprétation consiste à extraire du message oral des structures conceptuelles. Cette opé-

ration ne se résume pas forcément à une analyse de la transcription textuelle du message par une grammaire syntaxico-sémantique. Plusieurs considérations étayent cette proposition : d'une part, les règles d'interprétation peuvent être contextuelles ; d'autre part, dans le traitement de la parole spontanée, des parties entières du message peuvent être inutiles à la compréhension de celui-ci. L'opération d'interprétation conceptuelle peut être un succès, même si l'ensemble du message n'est pas complètement couvert par une grammaire. Enfin, la même séquence de mots peut être utile à la reconnaissance de plus d'un concept.

Plusieurs formalismes ont été proposés pour décrire des structures sémantiques, comme vu dans la partie 2.3.4. Ils sont essentiellement basés sur les concepts d'entités et de relations. En général, les concepts généraux représentant l'interprétation complète d'un message sont obtenus par des opérations de composition sur des concepts élémentaires. Ces concepts sont relativement indépendants du modèle sémantique global utilisé. Ils représentent à la fois les objets sémantiques manipulés par l'application, correspondant à des catégories d'entités nommées, telles que les dates, les prix, ou encore les noms propres (ville, hôtel, ...) ; mais aussi, les actes dialogiques. C'est sur ces concepts élémentaires que s'est focalisée la campagne d'évaluation MEDIA, qui consistait à évaluer les capacités d'interprétation de plusieurs systèmes sur un corpus de traces de dialogue homme-machine portant sur un serveur d'informations touristiques.

Ce chapitre présente mes travaux effectués au sein du LIA sur le corpus MEDIA et visant à proposer des méthodes d'analyse robustes permettant d'extraire une séquence de concepts élémentaires d'un message audio. Ces concepts sont les entités utilisées pour construire une interprétation sémantique complète des messages traités. La campagne MEDIA était structurée en deux phases : une phase d'évaluation de la compréhension *hors contexte* et une autre *en contexte*. Dans la première, les énoncés sont traités indépendamment les uns des autres, sans aucune information sur le dialogue en cours. Dans la deuxième, les concepts sont enrichis avec les informations contextuelles obtenues lors des précédents tours de dialogue. Nous nous focaliserons, dans cette étude, sur l'interprétation *hors contexte* des concepts élémentaires.

En complément de la campagne MEDIA, je présente également mes travaux effectués sur le corpus audio MEDIA. En utilisant, d'une part, le système de Reconnaissance Automatique de la Parole (RAP) SPEERAL [Nocera *et al.*, 2002] et, d'autre part, le module d'interprétation sémantique développé au LIA [Raymond *et al.*, 2006] nous montrons comment une approche de décodage *intégrée*, cherchant directement la meilleure séquence de concepts à partir d'un graphe de mots issus du module de RAP, surpasse l'approche séquentielle traditionnelle consistant à détecter les concepts uniquement dans la meilleure hypothèse de phrase sortie par le module de RAP.

Enfin, dans le cadre d'un décodage de la RAP, je montrerai l'influence du Taux d'Erreur Mot (ou Word Error Rate - WER) sur le Taux Erreur Conceptuel (ou Concept Error Rate - CER).

Ce chapitre est organisé comme suit : après avoir présenté les deux approches principales utilisées pour résoudre ce problème, celle basée sur une analyse séquentielle et celle qui envisage ce processus intégré, je présenterai le système de compréhension du

LIA basé sur une approche intégrée.

3.2 Méthode intégrée de décodage conceptuel

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, 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 doivent opérer sur des analyses partielles. En dehors de la traditionnelle opposition : *méthodes à base de connaissance a priori / méthodes à base d'apprentissage automatique*, les modèles développés pour répondre au problème du décodage conceptuel d'un flux de mots peuvent être vus selon deux perspectives :

- soit comme une conséquence d'un processus d'analyse syntaxico-sémantique 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 donnée en entrée (les mots) en une autre suite de symboles (les concepts).

Pour illustrer la première famille de méthodes, on peut se référer à la section 2.3.5 ; d'autres systèmes utilisent une analyse robuste incrémentale basée sur une étape de *chunking* comme dans [Antoine *et al.*, 2003]. Enfin, [Wang *et al.*, 2002] présente des grammaires hors-contextes syntaxico-sémantiques dérivées à partir de patrons génériques. Dans la campagne MEDIA, cette approche était représentée par les systèmes du VALORIA et du LORIA [Denis *et al.*, 2006].

La deuxième famille de méthodes se rapproche du cadre théorique utilisé dans les applications de Reconnaissance Automatique de la Parole (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 (des paramètres acoustiques dans le cadre de la RAP, des mots dans le cadre du décodage conceptuel) 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'observation acoustique. Cette approche, initiée par les travaux de [Levin et Pieraccini, 1995], se retrouve dans de nombreux systèmes de décodage conceptuel tels que [Miller *et al.*, 1994]. Dans la campagne MEDIA, les deux systèmes du LIMSI [Bonneau-Maynard et Lefevre, 2005] et celui du LIA mettaient en œuvre cette approche.

3.2.1 L'approche classique : la méthode séquentielle

La plupart des systèmes de compréhension de l'oral sont basés 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 dans la section 2.3.5. Cette approche est clairement sous-optimale, selon [Wang *et al.*, 2002], 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* : $\hat{W} = \underset{W}{argmax} P(W|A)$.

3.2.2 Une approche globale : la méthode intégrée

Afin de proposer une alternative à cette approche séquentielle, la méthode proposée intègre le processus de compréhension dans le processus de transcription de parole en remplaçant le modèle de langage par un modèle de décodage conceptuel. Pour permettre une intégration efficace, c'est la deuxième famille de méthodes de décodage conceptuel présentée précédemment qui a été choisie. Ce choix théorique est basé sur deux considérations :

- 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 *bi*-grammes ou *tri*-grammes), ne garantissant aucune cohérence au-delà d'une fenêtre de quelques mots ; pour cette raison, il est important de pouvoir garder des hypothèses multiples, sous la forme d'un graphe de mots ;
- même si de précédents travaux ont montré que des analyses syntaxiques peuvent être appliquées à des graphes de mots [Chappelier *et al.*, 1999, Roark, 2002], les disfluences et l'absence de structure dans le flux de parole rendent ces analyses difficiles, limitées le plus souvent à de multiples analyses partielles ; de fait, l'intérêt principal du processus d'analyse, l'obtention d'une analyse fonctionnelle complète de chaque composant du message, est rarement atteint.

3.3 Choix de la meilleure interprétation

Le modèle théorique utilisé est le suivant :

- soit A, la séquence d'observation acoustique représentant le message oral M ;
- soit $I = q_1 \dots q_n$, une interprétation de M représentée par la suite de concepts q_i ;
- soit $W = w_1 \dots w_k$, une suite de mots w_i supports d'une interprétation I ;

Le modèle d'interprétation 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 interprétation $\hat{I}$ du message M est fait avec la règle de décision du *Maximum a posteriori* selon la formule :

$$\begin{aligned}
\hat{I} = & \underset{I}{argmax}\ P(I|A) = \underset{I}{argmax} \sum_{W} P(W, I|A) \\
= & \underset{I}{argmax} \sum_{W} P(A|W, I)P(W, I) \\
= & \underset{I}{argmax} \sum_{W} P(A|W, I)P(W)P(I|W) \\
\approx & \underset{I,W}{argmax}\ P(A|W)P(W)P(I|W)
\end{aligned} \tag{3.1}$$

On remplace la probabilité $P(A|W, I)$ par $P(A|W)$ en considérant que la séquence d'observation A ne dépend que des mots W prononcés et non de l'interprétation I de ceux-ci.

Avec ce modèle, la meilleure interprétation $\hat{I}$ 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 $P(I|W)$ du modèle de décodage conceptuel représentant la probabilité d'une séquence de concepts I 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, permettent de faire porter un poids plus ou moins grand sur l'un ou l'autre des modèles. La formule 3.1 devient :

$$\hat{I} \approx \underset{I,W}{argmax} P(A|W)P(W)^{\alpha}P(I|W)^{\beta} \tag{3.2}$$

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

$$\begin{aligned}
\hat{I} \approx & \underset{I,W}{argmax}\ P(A|W)P(W)^{\alpha}P(W, I)^{\beta}P(W)^{-\beta} \\
\approx & \underset{I,W}{argmax}\ P(A|W)P(W)^{\alpha-\beta}P(W, I)^{\beta}
\end{aligned} \tag{3.3}$$

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 et sont présentées dans la section 3.10. L'estimation de la probabilité $P(W, I)$ est présentée dans la prochaine partie.

3.4 Modèle de langage conceptuel $P(W, I)$

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 $I = q_1 \dots q_n$.

Pour l'interprétation $I = q_1 \dots q_n$ et la chaîne de mots $W = w_1 \dots w_k$ nous avons :

$$P(W, I) = \sum_D P(W, D, I) \tag{3.4}$$

Ce problème s'apparente à un problème de segmentation de texte. 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. Pour cela, 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*.

Nous notons $(D, I) = (d_1, q_1)\ (d_2, q_2) \dots (d_n, q_n)$, et chaque couple (d_j, q_j) est remplacé par une séquence de d_j symboles t, tel que :

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

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 3.4 devient :

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

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

Par exemple, pour la séquence de concepts :

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

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, I)) :

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

d'où :

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

L'étiquetage correct étant :

$w_{1,4} = \texttt{oui à Marseille bon}$

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

L'estimation de $P(W, I)$ 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 :

$$P(w_{1,k}, t_{1,k}) = \prod_{i=1}^{n} P(t_i | t_{1,i-1}, w_{1,i-1}) P(w_i | t_{1,i}, w_{1,i-1}) \tag{3.5}$$

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

$$\begin{array}{rcl} P(t_i | t_{1,i-1}, w_{1,i-1}) & = & P(t_i | t_{i-2,i-1}, w_{i-2,i-1}) \\ P(w_i | t_{1,i-1}, w_{1,i-1}) & = & P(w_i | t_{i-2,i-1}, w_{i-2,i-1}) \end{array} \tag{3.6}$$

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 :

$$P(t_{1,n}, w_{1,n}) = \prod_{i=1}^{n} P(t_i | t_{i-2,i-1}, w_{i-2,i-1}) P(w_i | t_{i-2,i-1}, w_{i-2,i-1}) \tag{3.7}$$

Ce modèle de langage peut être appris directement sur un corpus d'apprentissage étiqueté avec les symboles t_i en utilisant le critère du maximum de vraisemblance. L'apprentissage et la représentation de ce modèle sont décrits dans la section 3.5.

3.5 D'un graphe de mot vers une liste structurée d'interprétations

Le module de compréhension de la parole du LIA, basé sur le modèle théorique de la partie précédante, prend en entrée un graphe de mots issu du module de RAP et produit une liste structurée d'interprétations du message. L'implémentation du modèle est détaillée dans cette section sur quelques données d'exemple. Toutes les opérations sur les FSMs sont réalisées avec les bibliothèques *FSM Library* et *GRM Library* de *AT&T Labs* [Mohri *et al.*, 2002, Allauzen *et al.*, 2004].

Il est à noter que nous allons utiliser un étiqueteur stochastique et des grammaires de concepts. L'utilisation seule d'un étiqueteur aurait pu suffire, mais notre but est de structurer l'espace de recherche créé par l'étiqueteur et ainsi permettre la réévaluation des scores des chemins du graphe. Les grammaires de concepts nous permettent d'effectuer cette tâche de structuration et de réévaluation.

3.5.1 Les grammaires de concepts

Chaque concept q_i d'une interprétation I peut être exprimé par une multitude de séquences de mots. Ces séquences de mots peuvent être représentées par une grammaire g_i, écrite manuellement ou apprise sur un corpus. Nous utilisons le formalisme des grammaires régulières que nous représentons sous la forme d'automates à états finis (*Finite State Machine* ou *FSM*) entièrement lexicalisés. Cette représentation, très restrictive, est justifiée par les deux remarques suivantes :

- Les concepts q_i représentent des unités très basiques (dates, noms propres, objets de la base de données, commande de dialogue,...) exprimées avec de courtes séquences de mots de tailles finies. Ils sont donc tous facilement décrits par des grammaires régulières.
- Une des fortes particularités du traitement automatique de l'oral est la contrainte posée par le lexique de reconnaissance utilisé par le module de RAP. En effet, si la capacité à traiter des mots inconnus est une fonctionnalité cruciale des systèmes traitant de l'écrit, cette capacité est inutile pour les systèmes traitant des sorties de modules de RAP : tous les mots susceptibles d'apparaître dans une transcription automatique sont dans le lexique de reconnaissance. Les mots inconnus du système prononcés par le locuteur vont soit être correctement détectés comme étant des mots inconnus et remplacés par un symbole `mot_inconnu`, soit générer des erreurs de substitution en étant remplacés par des mots du lexique de reconnaissance.

Ainsi, cette connaissance *a priori* de tous les mots pouvant exprimer un concept q_i nous permet de lexicaliser complètement les FSMs codant les grammaires g_i. L'ensemble des phrases supports L_I d'une interprétation $I = q_1 \dots q_n$ s'obtient donc directement en générant l'ensemble des phrases pouvant être reconnues par la grammaire $G_I = g_1\, g_2 \dots g_n$.

L'exemple suivant présente l'implémentation de cette méthode sur un petit corpus d'exemples.

Soit la liste de 8 extraits du corpus `MEDIA` avec les étiquettes conceptuelles posées manuellement par les annotateurs (nous n'avons gardé dans ce format que le type du concept et sa valeur, les informations sur les spécifieurs et le mode ont été supprimées) :

1. `je voudrais un <sem tag="OBD" value="hotel"> hôtel </sem>`
2. `<sem tag="LOC-VIL" value="paris"> à Paris </sem>`
3. `euh <sem tag="LOC-VIL" value="grenoble"> Grenoble </sem>`
4. `<sem tag="LOC-VIL" value="lyon"> dans Lyon </sem>`
5. `je prendrais l' <sem tag="NOM-HTL" value="mirande"> hôtel la Mirande </sem>`
6. `manger <sem tag="NOM-RES" value="david"> au David </sem>`
7. `<sem tag="NOM-HTL" value="georges_v"> le Georges-V </sem>`
8. `pour <sem tag="TMP-DTE" value="12/01/????"> le douze janvier </sem>`

L'étiquette de concept `OBD` correspond à la mention d'un objet de la base de données, `LOC-VIL` à la localisation d'une ville, `NOM-HTL` à une référence vers un hôtel de la base de données, `NOM-RES` à une référence vers un restaurant, `TMP-DTE` à l'expression d'une date.

À partir d'un tel corpus, sont extraites les séquences de mots exprimant chacun des concepts, telles qu'elles ont été définies par les annotateurs. Les entités numériques, les jours de la semaine, les mois et les noms propres sont généralisés avec des symboles non-terminaux tels que `XNUM`, `XMONTH` et `XNAME`. À chaque mot, est attachée l'étiquette du concept qu'il supporte ainsi que l'étiquette b s'il s'agit du premier mot exprimant le concept ou sinon l'étiquette i comme présenté dans la section 3.4.

1. `hôtel/OBD_b`
2. `à/LOC-VIL_b XNAME/LOC-VIL_i`
3. `XNAME/LOC-VIL_b`
4. `dans/LOC-VIL_b XNAME/LOC-VIL_i`
5. `hôtel/NOM-HTL_b la/NOM-HTL_i XNAME/NOM-HTL_i`
6. `au/NOM-RES_b XNAME/NOM-RES_i`
7. `le/NOM-HTL_b XNAME/NOM-HTL_i`
8. `le/TMP-DTE_b XNUM/TMP-DTE_i XMONTH/TMP-DTE_i`

Un automate transducteur est construit à partir d'un tel corpus : cet automate reconnaît en entrée des mots ou des symboles non-terminaux et produit en sortie des

étiquettes conceptuelles avec les informations *b* pour *begin* et *i* pour *inside*. Afin de modéliser les séquences de mots ne supportant aucun concept de l'application un modèle *mange-mot* ou *filler* est ajouté à ce transducteur, il prend en entrée tous les mots du lexique de RAP ainsi que les symboles non-terminaux et produit en sortie le symbole *null* (l'étiquette *O* ou *outside* du modèle *OIB*). Notons que tous les mots apparaissant dans le support d'un concept sont aussi dans le *mange-mot*, car c'est le contexte d'occurrence d'une séquence de mots qui permet de décider s'il s'agit ou non d'un concept de l'application.

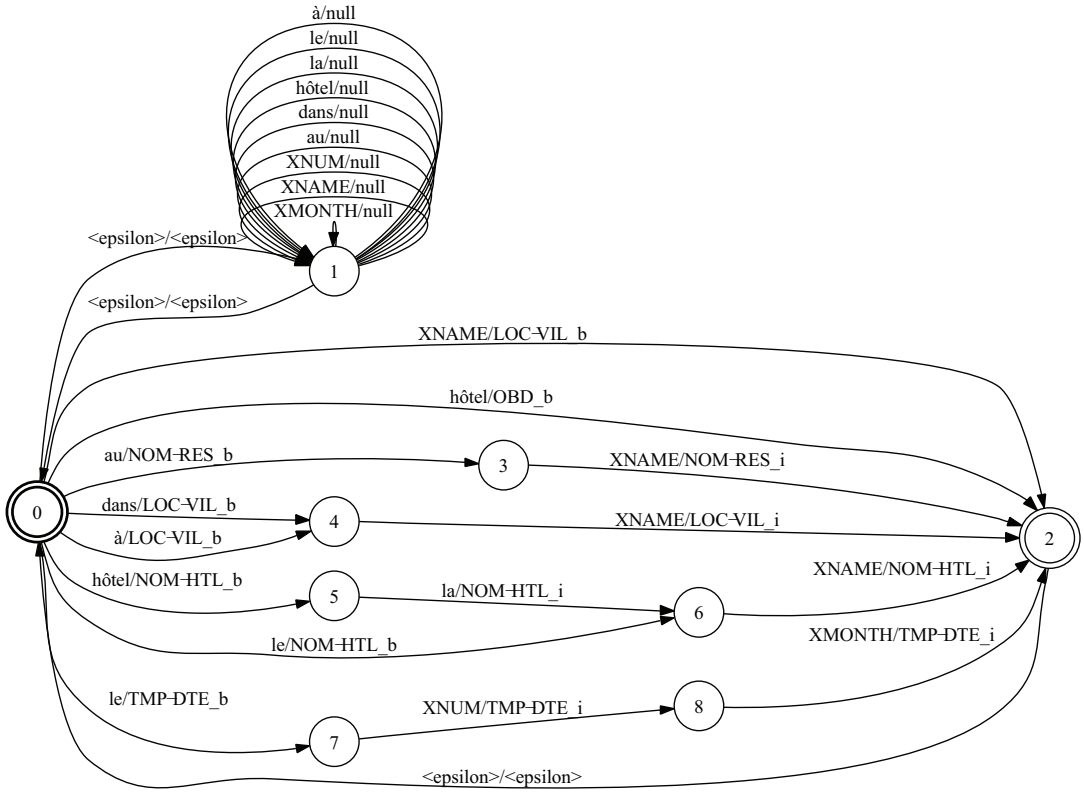

FIG. *3.1:* *Transducteur des grammaires de concept : $T_{concept}$*

Un exemple d'un tel transducteur, appelé $T_{concept}$, est donné dans la figure 3.1. Comme ce transducteur n'est pas évalué, il est très facile d'y ajouter des grammaires manuelles représentant une connaissance *a priori* sur certains concepts de l'application. Par exemple, pour l'application `MEDIA`, nous avons ajouté des grammaires manuelles codant les dates, les expressions numériques, telles que les prix et les entités de la base de données comme les villes, les noms d'hôtels et de restaurants.

3.6 D'un graphe de mots vers un graphe de concepts

Les chemins W du graphe de mots G issu du module de RAP sont évalués avec les probabilités $P(A|W)P(W)^{\alpha-\beta}$ des modèles acoustiques et du modèle de langage qui sont présentés à la section 3.10. La première opération réalisée par le module d'interprétation consiste à le structurer en déterminant toutes les séquences $I = q_1 \dots q_n$ pouvant être reconnues dans ce graphe. Cette opération est effectuée en deux temps : tout d'abord, transformer le graphe G en transducteur T_G en faisant émettre à chaque symbole w, soit lui-même, soit un symbole non-terminal, tel que `XNAME`. Ensuite, ce transducteur est composé avec le transducteur $T_{concept}$ afin d'obtenir l'ensemble des séquences de concepts contenues dans le graphe.

Un exemple de graphe G avec le transducteur correspondant T_G est donné dans la figure 3.2. Les valeurs portées par les transitions du graphe et représentant les probabilités des modèles de RAP ne sont pas indiquées dans la figure. Le résultat de l'intersection $T_G \cap T_{concept}$ est présenté dans la figure 3.3.

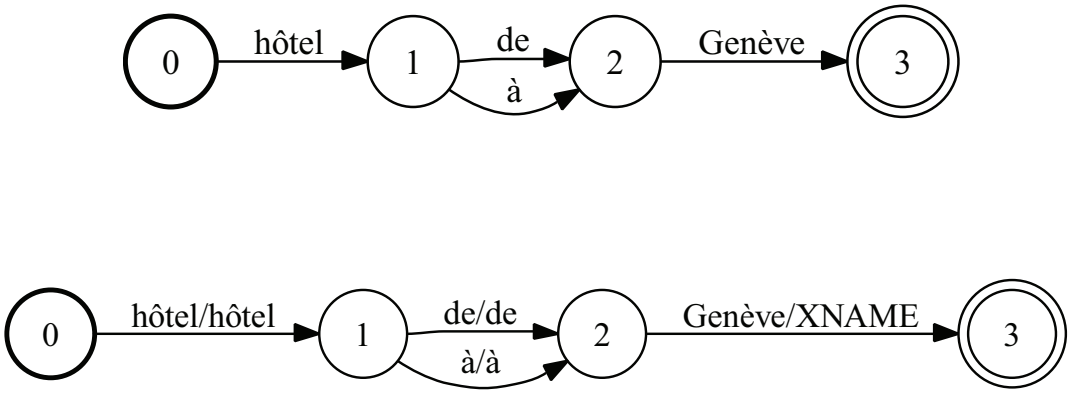

FIG. 3.2: *Graphe de mot G issu du module de RAP avec sa généralisation T_G*

Jusqu'à cette étape, les seules valeurs portées par les chemins du graphe sont les probabilités $P(A|W)P(W)^{\alpha-\beta}$. Chaque chemin va ensuite être réévalué en lui appliquant le modèle conceptuel $P(W, I)$ présenté dans la section 3.4. Ce modèle est appris sur un corpus contenant des transcriptions de dialogues avec les étiquettes conceptuelles présentées auparavant. Sur l'exemple présenté en section 3.5.1, le corpus d'apprentissage du modèle $P(W, I)$ devient :

1. `je/null voudrais/null un/null hôtel/OBD_b`
2. `à/LOC-VIL_b XNAME/LOC-VIL_i`
3. `euh/null XNAME/LOC-VIL_b`
4. `dans/LOC-VIL_b XNAME/LOC-VIL_i`

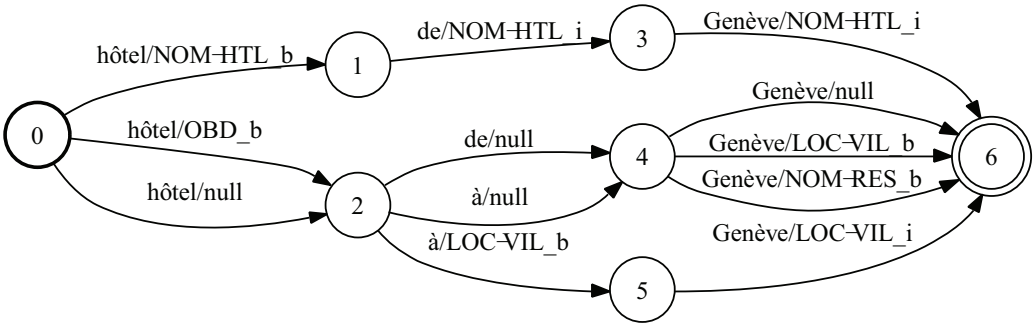

FIG. 3.3: *Graphe de mot enrichi avec les séquences de concepts : $T_G \cap T_{concept}$*

5. `je/null prendrais/null l'/null hôtel/NOM-HTL_b la/NOM-HTL_i XNAME/NOM-HTL_i`
6. `manger/null au/NOM-RES_b XNAME/NOM-RES_i`
7. `le/NOM-HTL_b XNAME/NOM-HTL_i`
8. `pour/null le/TMP-DTE_b XNUM/TMP-DTE_i XMONTH/TMP-DTE_i`

Ce modèle est représenté lui aussi sous la forme d'un automate, appelé $LM_{concept}$, grâce à la bibliothèque *GRM Library* de *AT&T Labs* [Allauzen *et al.*, 2004]. L'application de ce modèle conceptuel au transducteur $T_G \cap T_{concept}$ se fait simplement par l'opération d'intersection (ou *composition*) avec l'automate $LM_{concept}$.

À la suite de cette intersection, les chemins du transducteur $T_G \cap T_{concept} \cap LM_{concept}$ sont évalués selon la formule 3.3. En projetant ce transducteur sur les symboles de sorties et en ne gardant que les symboles marquant le début d'un concept (avec l'extension `_b`), nous obtenons un graphe conceptuel de toutes les interprétations possibles. Sur l'exemple précédent, les interprétations possibles I sont :

1. `NOM-HTL_b`
2. `OBD_b LOC-VIL_b`
3. `OBD_b NOM-RES_b`
4. `OBD_b`
5. `LOC-VIL_b`
6. `NOM-RES_b`

Ces hypothèses constituent le premier niveau dans la liste structurée d'interprétations que le module de compréhension produit. Le deuxième niveau est constitué par la liste des meilleures valeurs pour chaque concept des différentes interprétations.

3.7 D'un graphe de concepts vers un graphe de valeurs

Nous disposons à présent d'une liste structurée d'interprétations. Il nous faut à présent extraire les informations sur les valeurs contenues dans cette liste. Pour cela, nous devons conserver les séquences de mots associées aux solutions. Grâce à ces séquences de mots associées aux concepts trouvés, nous allons pouvoir extraire les valeurs associées aux concepts trouvés. Nous utiliserons les bibliothèques *FSM Library* et *GRM Library* de *AT&T Labs*, pour modéliser les grammaires de valeurs sous forme de FSM.

3.7.1 Les grammaires de valeurs

Chaque valeur v d'un concept Q est composée de multiples séquences de mots. Ces séquences de mots sont modélisées par des grammaires g_v qui sont soit apprises sur un corpus, soit écrites manuellement. Nous représentons ces grammaires sous forme d'automates à états finis, ou *FSM*. Ces grammaires se prêtent à ce formalisme, car :

- elles sont composées de séquences de mots simples avec un lexique de taille fini ;
- elles sont un nombre fini car elles ont été créées à partir d'une liste de valeurs normalisée ;
- elles sont de tailles finies et ne sont pas redondantes.

Nous ferons l'union des grammaires régulières g_v (sous forme de FSM) dans un transducteur de valeurs T_v.

Nous allons associer des valeurs aux concepts que nous avons identifiés. Par exemple, préciser le nom de la ville lorsque nous reconnaissons une ville. Pour ce faire, reprenons notre jeu d'exemple précédent d'où nous allons noter les valeurs associées.

1. `hôtel_OBD_b_nom-hotel`
2. `à_LOC_VIL_b_null Paris_LOC_VIL_i_paris`
3. `Grenoble_LOC_VIL_b_grenoble`
4. `dans_LOC_VIL_b_null Grenoble_LOC_VIL_i_grenoble`
5. `hôtel_NOM_HTL_b_null la_NOM_HTL_i_null XNAME_NOM_HTL_i_mirande`
6. `au_NOM_RES_b_null David_NOM_RES_i_chez_david`
7. `le_NOM_HTL_b georges_NOM_HTL_i_null V_NOM_HTL_i_georges_v`
8. `le_TMP_DTE_b_null douze_TMP_DTE_i_12D janvier_TMP_DTE_i_01M`

De même que pour les grammaires conceptuelles, les grammaires de valeurs g_v sont induites à partir d'exemples. Nous faisons l'union de ces grammaires dans un seul transducteur noté T_v. Contrairement au transducteur $T_{concept}$, T_v ne possède pas de chemin permettant une attribution de valeur *nulle* ou *inconnue* à un concept identifié. Si le chemin n'a pas de valeur, il est tout simplement supprimé du graphe de solution.

$$G_{Valeurs} = G_{Concepts} \cap T_v \tag{3.8}$$

Ce transducteur (figure 3.4) de valeur agit donc comme un filtre sur le graphe de solutions. Les règles ci-dessus sont induites par un corpus d'apprentissage, mais nous

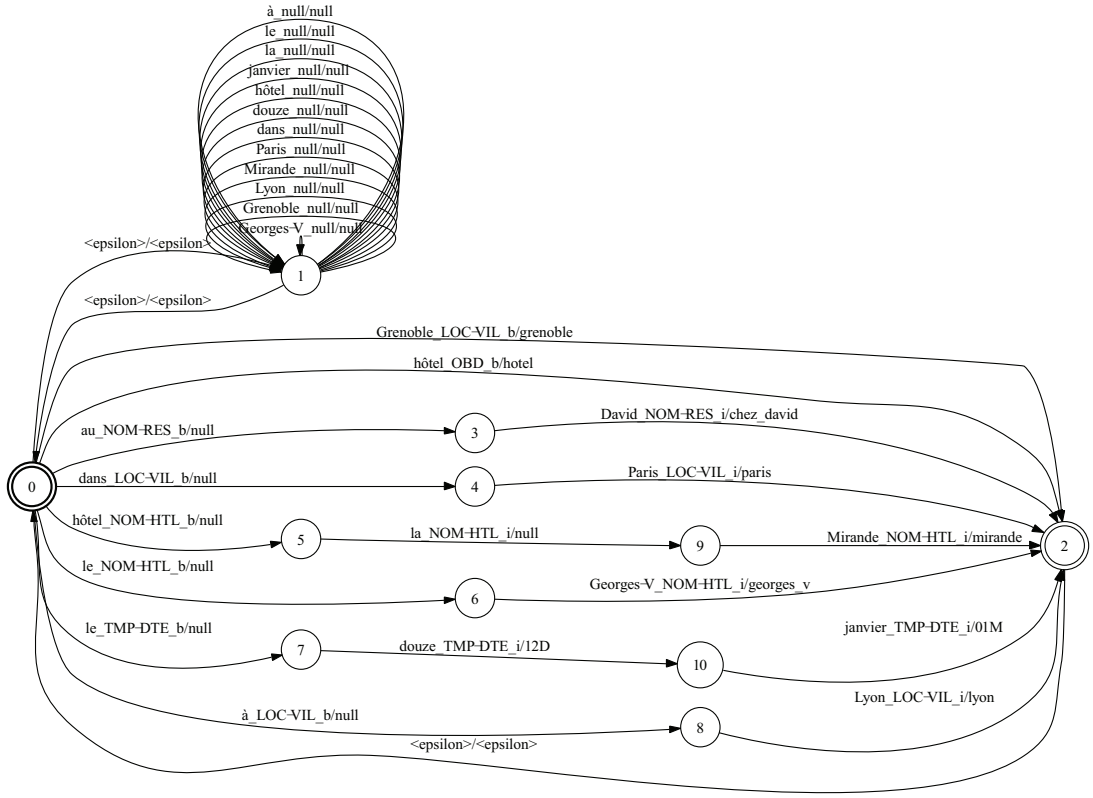

FIG. *3.4:* *Transducteur des grammaires de valeur*

utilisons aussi des règles de grammaire créées à partir de bases de connaissance. Elles ne sont pas spécifiques au corpus `MEDIA`, elles se présentent sous forme de dates, de chiffre, etc. Les automates créés de cette manière sont utilisés pour compléter les automates obtenus par apprentissage automatique.

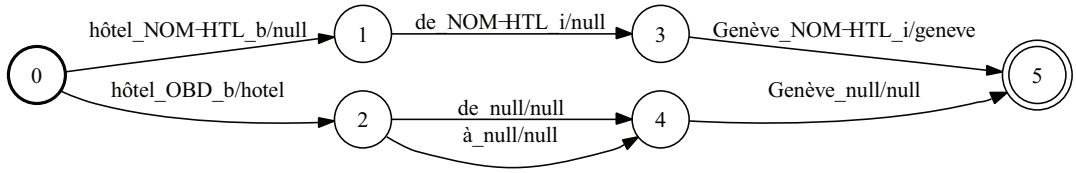

FIG. *3.5:* *Graphe de mot/concept enrichi des valeurs*

L'intersection entre le graphe de solutions enrichi de concepts et le transducteur T_v donne un graphe de solutions figure 3.5.

Sachant que les opérations de transduction n'ont pas affecté les scores de chemins, nous pouvons maintenant obtenir les n-meilleures interprétations contenues dans le graphe $G_{Valeurs}$ (figure 3.5). De plus, grâce aux symboles d'entrées de ce graphe, nous

pouvons aussi obtenir la chaîne de mots W^V associée à la meilleure interprétation conceptuelle avec les valeurs V.

3.8 La liste de N-meilleures solutions structurées

Pour exploiter le graphe de solutions, nous pouvons organiser les solutions en créant une liste structurée de solution. Le principe consiste à sélectionner la meilleure interprétation sans tenir compte des différentes valeurs qui sont attribuées aux chaînes de concepts. Nous rassemblons les meilleures solutions associées aux valeurs pour chaque interprétation. Nous obtenons une liste structurée d'interprétations et de valeurs comme présenté dans le tableau 3.1.

Chaîne de mots	hôtel de Genève	
Interprétation 1	nom-hotel	
valeurs 1	genève	
Interprétation 2	objetBD	null
valeurs 1	hotel	null

TAB. *3.1:* *Liste de solutions structurées*

3.9 Architecture globale du système de décodage

Toutes les opérations qui viennent d'être décrites peuvent être résumées dans un schéma présentant l'architecture du système, figure 3.6.

Nous pouvons voir, sur ce schéma, le passage d'un graphe de mots à un graphe de solutions, enrichi de concepts et composé avec un modèle de langage conceptuel. Enfin, le graphe de solutions de concepts est composé avec le transducteur des valeurs, pour donner un graphe de solutions avec une liste d'interprétations chacune associée à une liste de valeurs.

3.10 Développement d'un système de RAP sur le corpus MEDIA

3.10.1 Apprentissage des modèles

Le décodeur SPEERAL [Nocera *et al.*, 2002] a été utilisé pour transcrire les messages du corpus MEDIA. Ces messages sont enregistrés avec un système de Magicien d'Oz, c'est-à-dire une similation des réactions d'une machine par un intervenant humain. Les utilisateurs ont effectué leurs appels depuis leur téléphone (fixe ou cellulaire) et la qualité des enregistrements est variable. Les modèles acoustiques téléphoniques utilisés sont ceux développés lors de la campagne d'évaluation ESTER sur la transcription de

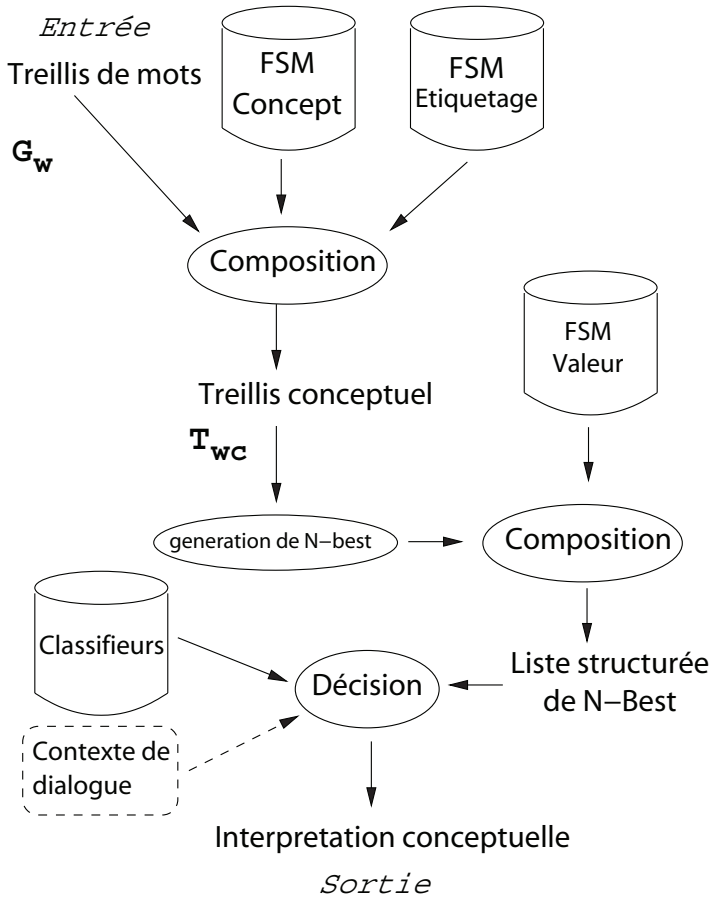

FIG. 3.6: *Stratégie de compréhension de la parole du LIA*

données radiophoniques, ils ont ensuite été adaptés sur les 720 dialogues des lots 1,2,3,4 du corpus MEDIA par une adaptation de type MAP[1].

Le modèle de langage a été appris sur un corpus extrait des transcriptions manuelles des lots 1,2,3,4. Ce corpus contient un ensemble de 226K mots. Un lexique de 2 028 mots a été défini sur ce corpus, il a été phonétisé avec l'outil LIA_PHON[2]. Sur le corpus de test utilisé (lot *Test à blanc* du corpus MEDIA), le taux de mots hors-vocabulaire du lexique choisi est de 1,6%. La perplexité est de 26,5.

Le taux d'erreur mot (ou *Word Error Rate WER*) de la transcription automatique du

[1]MAP ou Maximum A Posteriori : on recherche la situation ou la valeur initiale des paramètres la plus probable connaissant la situation ou la valeur finale

[2]téléchargeable à l'adresse :
http ://www.lia.univ-avignon.fr/chercheurs/bechet/

lot *Test* avec les modèles présentés est de 32,2%.

3.10.2 Variations du WER

L'approche intégrée de décodage conceptuel défendue dans cette partie nécessite le traitement de graphes de mots issus du module de RAP. Les graphes nous permettent également de faire varier le WER de la meilleure hypothèse produite par le module de RAP. En effet, un but de cette section est d'étudier la corrélation entre le taux d'erreur sur les mots et celui sur les concepts. Il est donc intéressant de produire des sorties multiples.

Nous disposons de plusieurs méthodes afin de créer ces différentes sorties. La première que nous avons choisie est l'introdution de données de tests dans l'apprentissage, la seconde est une combinaison de systèmes.

Introduction des données de test

L'introduction de données de test dans l'apprentissage est obtenue par la méthode suivante :

- tout d'abord des graphes de mots sont générés par SPEERAL sur le corpus de test avec les modèles présentés précédemment ; en prenant la meilleure séquence de mots dans ces graphes, nous obtenons les hypothèses de référence (ou *baseline*) avec un WER moyen de 32,2% ;
- un nouveau modèle de langage est alors appris, cette fois sur le corpus de test ;
- ce modèle est appliqué aux graphes de mots préalablement produits, après une interpolation avec le modèle *baseline* appris sur le corpus d'apprentissage ;
- en faisant varier le coefficient d'interpolation, on peut faire varier le taux d'erreur mots.

Avec cette méthode nous avons obtenu 4 décodages différents de notre corpus de test obtenus avec 4 valeurs différentes du coefficient d'interpolation (0,0 0,5 0,8 et 1.0). Ces décodages sont représentés par 4 séries de graphes de mots $G_{0,0}$, $G_{0,5}$, $G_{0,8}$ et $G_{1,0}$. Les graphes $G_{0,0}$ correspondent au décodage *baseline* où aucune donnée de test n'est intégrée dans l'apprentissage. Les taux d'erreurs mots des meilleures hypothèses de ces graphes sont présentés dans le tableau 3.2.

Graphes	$G_{0,0}$	$G_{0,5}$	$G_{0,8}$	$G_{1,0}$
WER	32,2	27,2	24,1	18,5

TAB. 3.2: *WER des graphes correspondants aux différentes interpolations*

Même si l'introduction de données de tests dans l'apprentissage génère forcément un biais, il est réduit du fait que cette introduction n'intervient que dans la deuxième passe de l'étape de reconnaissance : les erreurs et confusions acoustiques produites par le modèle *baseline* sont toujours présentes. Cependant, ce sont bien évidemment les résultats obtenus sur les graphes $G_{0,0}$ qui sont les plus réalistes, puisqu'ils sont produits

sans introduction des données de tests. Les autres graphes ne servent qu'à observer la corrélation taux d'erreur mots et taux d'erreur concepts dans la section 4.3.1.

Combinaisons de systèmes de RAP

Grâce aux travaux réalisés dans le projet DIVINES [M.Benzequida *et al.*, 2006], le LIA a proposé des méthodes de combinaisons de systèmes [Barrault, 2008]. Dans ce but, il a été proposé de combiner les probabilités *a posteriori* des états de HMMs au niveau de la trame, comme présenté dans la figure 3.7. Une manière de combiner les probabilités *a posteriori* a été envisagée : la combinaison log-linéaire [Barrault *et al.*, 2008].

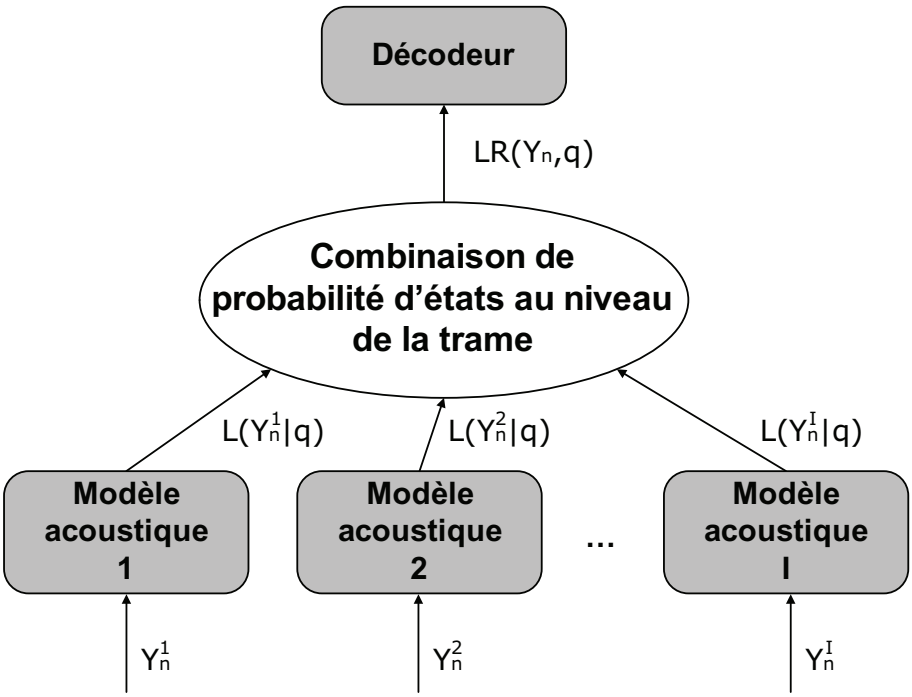

FIG. 3.7: *Architecture pour la combinaison trame à trame*

La combinaison log-linéaire (notée LLC) des probabilités postérieures est effectuée en supposant que chaque espace acoustique fournit une quantité d'information proportionnelle aux probabilités *a posteriori*. Les modèles acoustiques ont des GMMs qui modélisent des fonctions de densités de probabilités associées à chaque état. Les mixtures de gaussiennes permettent de calculer les vraisemblances $L(Y_n^i|q)$ qu'un état q ait généré une trame Y_n^i. Les vraissemblances $L(Y_n^i|q)$ sont calculées de manière simultanée pour chaque jeu de paramètres. Puis, pour chaque trame, un taux de vraissemblance $LR(Y_n, q)$ est calculé en utilisant la combinaison log-linéaire.

Les sorties sont basées sur cette combinaison des probabilités *a posteriori*. Nous uti-

lisons pour ce faire trois systèmes avec des jeux de paramètres différents et une combinaison de ces trois systèmes :

- *PLP* : le système à base d'analyse prédictive linéaire perpétuelle (*Perceptual Linear Prediction*) repose sur un modèle de perception de la parole. Les PLP sont basés sur le spectre à court terme du signal de parole. Cela signifie que le signal est analysé sur une fenêtre glissante de courte durée qui, en général, est de longueur 10 à 30 ms que l'on décale de 10 ms pour chaque trame [Hermansky, 1990] ;
- *RPLP* : système basé sur le *PLP* avec un filtre RASTA (RelAtive SpecTrAl) afin d'augmenter la robustesse des paramètres PLP. C'est une façon de simuler l'insensibilité de l'appareil auditif humain aux stimuli à variation temporelle lente. En pratique, RASTA effectue un filtrage passe-bande sur le spectre logarithmique ou sur le spectre compressé par une fonction non linéaire. L'idée principale est de supprimer les facteurs constants dans chaque composante du spectre à court terme avant l'estimation du modèle d'ensemble.
- *MRA* : système à Analyse à Résolution Multiple (*Multiple Resolution Analysis*). Il effectue une analyse en ondelettes d'une fenêtre de signal audio. Cela consiste à faire passer le signal dans un arbre de filtres passe-bas et passe-haut, à la sortie desquels l'énergie à court terme est calculée. À chaque niveau de l'arbre, le signal est entièrement décrit, mais dans une résolution fréquentielle et temporelle différente [Gemello *et al.*, 2006].
- *LLC* : la combinaison des trois systèmes décrits dans cette section.

Avec ces méthodes, nous avons obtenu 4 décodages différents de notre corpus de test. Ces décodages sont représentés par 4 séries de graphes de mots G_{PLP}, G_{RPLP}, G_{MRA} et G_{LLC}. Les graphes G_{PLP} correspondent au décodage *baseline* qui avait déjà été utilisé pour d'autres expériences [Servan *et al.*, 2006]. Les taux d'erreurs mots des meilleures hypothèses de ces graphes sont présentés dans le tableau 3.3.

Graphes	G_{PLP}	G_{RPLP}	G_{MRA}	G_{LLC}
WER	32,0	33,0	33,0	27,6

TAB. 3.3: *WER correspondant aux différents graphes produits par les systèmes associés.*

Nous avons décidé d'exploiter cet espace de recherche en modifiant les données du modèle acoustique, de manière à obtenir un meilleur taux d'erreur mot (ou *WER*) pour différents graphes.

Enfin, grâce à ces différents graphes, nous pourrons observer la corrélation taux d'erreur mots et taux d'erreur concepts.

3.10.3 Synthèse

Nous disposons, dès à présent, de plusieurs graphes de mots issus du module de reconnaissance de la parole. Ces graphes vont nous permettre d'effectuer des expériences afin d'établir s'il existe une correlation entre le taux d'erreur mot (*WER*) et le taux d'erreur conceptuel (*CER*). Ces expériences sont décrites dans le chapitre suivant, dans la partie 4.2.

3.11 Conclusion

Nous avons présenté notre approche du décodage conceptuel, ainsi que notre système. Notre système nous permet aussi bien d'effectuer un décodage séquentiel qu'un décodage intégré. Grâce à la méthode intégrée, nous allons pouvoir exploiter les graphes issus de la RAP et éventuellement corriger, potentiellement, la transcription. Le chapitre suivant présente les différentes expériences faites avec notre système.

Chapitre 4

Expériences de décodage de la parole spontanée

Sommaire

Ce chapitre sera consacré aux résultats des expériences menées avec le système de compréhension décrit dans le chapitre précédent. Il présente mes travaux effectués sur le corpus MEDIA utilisant une approche intégrée. Seront exposés les résultats de mes diverses expériences, tout d'abord, celles effectués en respectant le protocole d'évaluation de la campagne MEDIA. Ensuite j'exposerai les travaux sur l'intégration de connaissances *a priori* dans mes modèles stochastiques conceptuels (section 4.4), puis, celles faites sur la parole dans la section 4.3. Enfin, nous avons procédé à une analyse d'erreur dont le compte rendu se trouve en section 4.5.

	Full		Relax	
	4 modes	2 modes	4 modes	2 modes
LIA	41.3	36.4	29.8	24.1
LIMSI-1	**29.0**	23.8	**27.0**	21.6
LIMSI-2	30.3	**23.2**	27.2	**19.6**
LORIA	36.3	28.9	32.3	24.6
VALORIA	37.8	30.6	35.1	27.6

TAB. 4.1: *Tableau présentant les résultats de la campagne d'évaluation MEDIA des différents participants. Les chiffres sont des taux d'erreur conceptuels (CER)*

Type d'éval.		Tx Del.	Tx Sub.	Tx Ins.	CER	ok
full	4 modes	5.94	19.32	6.39	31.66	74.72
	2 modes	6.04	14.05	6.48	26.58	79.89
Relax	4 modes	5.99	14.22	6.43	26.66	79.77
	2 modes	6.09	8.56	6.53	21.18	85.34

TAB. 4.2: *Résultats avec le protocole MEDIA*

4.1 Expériences respectant le protocole d'évaluation MEDIA

Nous avons rassemblé dans le tableau 4.2 les résultats de tous ces modes avec la dernière version du système. Les expériences ont été faites sur la transcription manuelle du corpus de test officiel fourni pour la campagne d'évaluation. Pour mémoire, le tableau 4.1 contient les résultats de l'évaluation de la campagne MEDIA. Nous pouvons observer que nous sommes à présent au même niveau que les systèmes présents lors de l'évaluation hors-contexte [Bonneau-Maynard *et al.*, 2006].

Le système a évolué durant tout cette thèse par la correction de bogues et les améliorations apportées, ce qui explique la différence de résultats. Les principales améliorations sont celles des ressources (notamment des grammaires de détection de concepts et celles de valeurs), en ajoutant des grammaires de détection manuelles et en corrigeant le corpus d'apprentissage. Par exemple, la valeur erronée « 12 » était associée au concept « localisation-ville » et aux mots « à Paris », la valeur corrigée est « paris ». Enfin à force d'utiliser le corpus de test, il devient peu à peu un corpus de développement. Nous pouvons ainsi mesurer quelle a été l'évolution du système, entre la fin de la campagne et aujourd'hui.

Les résultats du tableau 4.2 ont été obtenus grâce à l'outil MEDIAVAL (section 2.4.8) créé spécialement pour l'évaluation.

4.2 Expériences relatives au traitement d'hypothèses multiples de la RAP

La série d'expériences présentée dans cette section a été menée sur le corpus MEDIA en considérant les 83 attributs présentés dans la section 2.4.2. Le mode et les 19 spécifieurs ne sont pas pris en compte ici, ils sont traités dans notre système par le module d'interprétation d'un énoncé en contexte et ne relèvent pas du processus de décodage conceptuel présenté dans la section 3.5. Les modalités de l'évaluation sont comparables à celles de l'évaluation MEDIA en mode *relax*, avec les valeurs et sans la modalité.

L'expérience suivante concerne l'évaluation des listes de n-meilleures hypothèses produites par les différentes méthodes testées. Ces listes sont particulièrement intéressantes dans le cadre d'un dialogue, car il est possible de fournir au gestionnaire de dialogue, non pas une hypothèse unique, mais plusieurs hypothèses, que le contexte du dialogue peut aider à filtrer. Une mesure communément utilisée, pour mesurer le potentiel d'un graphe ou d'une liste d'hypothèses, est la mesure *Oracle*.

Cette mesure consiste à sélectionner dans un ensemble d'hypothèses, celle qui a le plus petit taux d'erreurs. Elle constitue le taux d'erreur minimal que ferait un système qui prendrait toujours la bonne décision sur le filtrage d'une liste d'hypothèses. Trois listes d'hypothèses ont été produites à partir des graphes G_{PLP} (graphes *baseline* n'incluant pas de corpus de test dans l'apprentissage des modèles), leurs évaluations sont présentées sur la figure 4.1 :

- *séquentielle* : cette liste est obtenue en énumérant la liste des n-meilleures hypothèses obtenues avec la méthode séquentielle ; nous conservons la meilleure chaîne de mots issue de la RAP, à partir de laquelle nous allons décoder, afin d'obtenir les n-meilleures hypothèses conceptuelles ;
- *intégrée* : cette fois les n-meilleures hypothèses sont les n-meilleurs chemins dans le transducteur généré avec la méthode intégrée ; nous utilisons, non plus la meilleure interprétation issue de la RAP, mais le graphe de complet d'hypothèses issu du module de RAP. Ce graphe d'hypothèses est interprété dans le module de compréhension, qui génère les n-meilleures hypothèses d'interprétation.

Comme le montre la figure 4.1, la méthode intégrée permet d'éviter le principal inconvénient des listes de n-meilleures hypothèses produites à partir de la meilleure hypothèse du module de RAP : la génération d'hypothèses qui ne diffèrent que par des concepts non signifiants du point de vue de l'interprétation du message.

En produisant une liste par chaînes de concepts et de valeurs, on obtient un résumé de toutes les interprétations possibles contenues dans le graphe. Par exemple, en ne gardant que les 3 meilleures hypothèses de la méthode intégrée, on obtient le même CER qu'avec la liste 7 meilleures hypothèses de la méthode séquentielle.

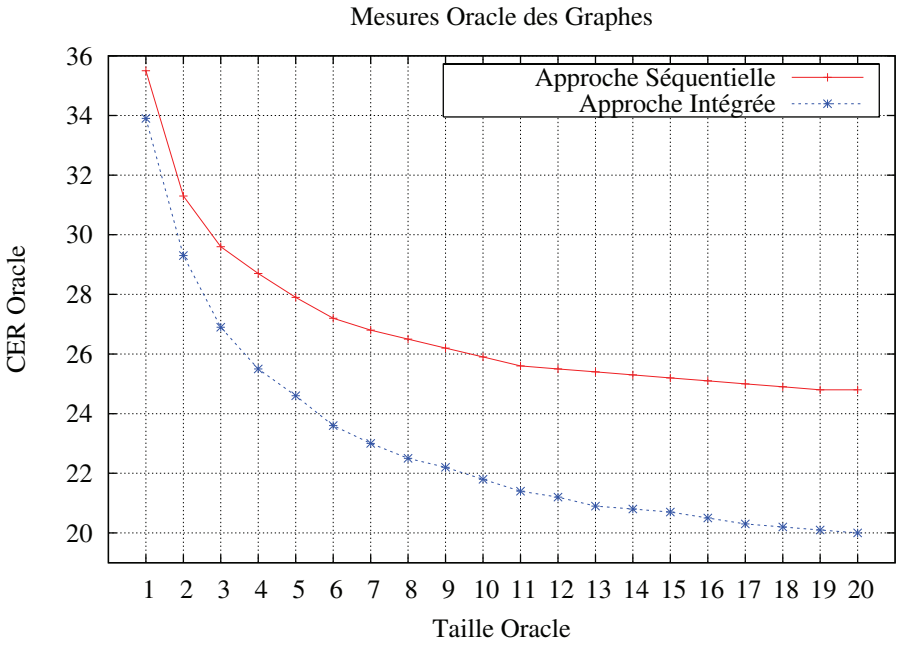

FIG. 4.1: *Évolution du CER Oracle en fonction des tailles des listes de n-meilleures hypothèses pour deux méthodes : méthode séquentielle et méthode intégrée*

4.3 Influence du taux d'erreur mot (WER) sur le taux d'erreur conceptuel (CER)

L'approche intégrée de décodage conceptuel nécessite le traitement de graphes de mots issus du module de RAP. Ces graphes sont produits par le décodeur SPEERAL, toutes les opérations sur les graphes sont ensuite effectuées avec l'ensemble d'outils de manipulation d'automates *AT&T FSM/GRM Library* [Mohri *et al.*, 2002]. Ces graphes nous permettent également de faire varier le WER de la meilleure hypothèse produite par le module de RAP. En effet, un but de cette méthode est d'étudier la corrélation entre le taux d'erreur sur les mots et celui sur les concepts. Il est donc intéressant de produire des sorties multiples.

Les expériences de cette partie présentent l'influence du *WER* sur le taux d'erreurs sur les paires attribut/valeur (appelé le *Concept Error Rate* ou *CER*). En réutilisant les graphes acoustiques modifiés par les coefficients d'interpolations, nous avons voulu mesurer l'impact du *WER* sur le *CER*.

Les expériences présentées ont été menées sur le corpus `MEDIA` en considérant les 83 attributs présentés à la section 2.4.2. Les modalités d'évaluation des expériences sont les mêmes que dans la section 4.2.

Les performances sont mesurées par rapport au taux d'erreurs sur les paires attribut/valeur (appelé le *Concept Error Rate* ou *CER*). Un concept détecté est considéré comme correct uniquement si l'attribut du concept ainsi que sa valeur normalisée sont corrects d'après la référence.

4.3.1 Expériences sur la RAP avec introduction de données de test

Le tableau 4.3 présente les résultats des deux approches, *séquentielle* et *intégrée*, sur plusieurs graphes de mots que nous avons présentés dans la partie 3.10.2. Nous observons que dans tous les cas l'approche intégrée surpasse l'approche séquentielle de manière significative, sauf bien sûr pour les transcriptions manuelles où le graphe de mots est réduit à une seule chaîne.

Graphe	$G_{0,0}$	$G_{0,5}$	$G_{0,8}$	$G_{1,0}$	Ref.
WER	32,2	27,2	24,1	18,5	0
CER (Seq.)	44,8	41,2	39,3	36,5	20,9
CER (Int.)	40,8	38,5	37,7	34,2	20,9

TAB. 4.3: *WER et CER sur différents graphes avec l'approche séquentielle (Seq.) et l'approche intégrée (Int.). La colonne Ref. correspond au traitement de la transcription manuelle du corpus de test*

Un autre enseignement intéressant de ces expériences est la corrélation entre le taux d'erreur sur les mots (WER) et celui sur les concepts (CER). La figure 4.2 l'illustre en montrant une relation linéaire entre ces deux quantités.

4.3.2 Expériences avec la combinaison de systèmes de RAP

Le tableau 4.4 présente les résultats de l'approche *séquentielle* sur plusieurs graphes de mots présentés dans la partie 3.10.2.

Malheureusement, il n'a pas été possible pour le système de RAP de fournir un graphe complet de solutions exploitable pour le système, afin d'expoiter la méthode *intégrée*. Nous avons donc seulement la meilleure hypothèse du module de reconnaissance de la parole, sur laquelle le module de compréhension a effectué un décodage conceptuel. Néanmoins, nous pouvons très nettement voir l'influence du WER sur le CER.

Graphe	G_{RPLP}	G_{MRA}	G_{PLP}	G_{LLC}	Ref.
WER	33,0	33,0	32,0	27,6	0
CER Seq.	37,1	37,0	35,1	32,4	18,5

TAB. 4.4: *WER et CER sur différents graphes avec l'approche séquentielle (Seq.). La colonne Ref. correspond au traitement de la transcription manuelle du corpus de test*

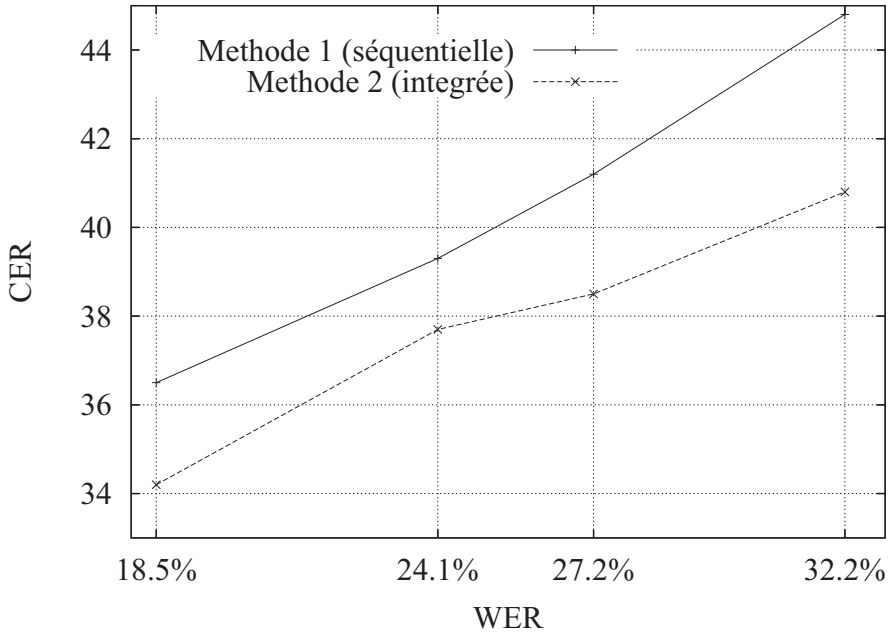

FIG. 4.2: *Évolution du CER en fonction du WER*

Au vu des résultats du tableau 4.4, il apparaît que la relation entre le *CER* et le *WER* est linéairement dépendante.

Enfin, nous observons, sur la figure 4.4, une évolution semblable à la figure 4.1 concernant le *CER* jusqu'à la taille oracle 20. Dans cette dernière figure, la courbe associée à la combinaison LLC est très nettement en dessous des autres, ayant un CER de 22 lorsque la taille de l'oracle est à 20.

On remarque qu'avec une méthode adéquate permettant de sélectionner la bonne hypothèse dans la liste des n-meilleures hypothèses conceptuelles, on pourrait diminuer le CER de plus de 30% relativement au résultat n'utilisant que la meilleure hypothèse.

Il convient de relever les différences entre les résultats des deux séries d'expériences présentées dans les tableaux 4.3 et 4.4. En effet, il peut paraître surprenant, voir incongru que malgré la différence entre les deux *WER* (soit moins de 1 point) on obtienne des différences de résultat si importantes sur le CER. Cela est dû aux différentes versions du système. En effet, les premières expériences ont été faites avec la version du système utilisée pour la campagne `MEDIA`, tandis que la deuxième série de résultats a été faite avec la version actuelle du système. Nous pouvons ainsi mesurer quelle a été l'évolution du système, entre la fin de la campagne et aujourd'hui.

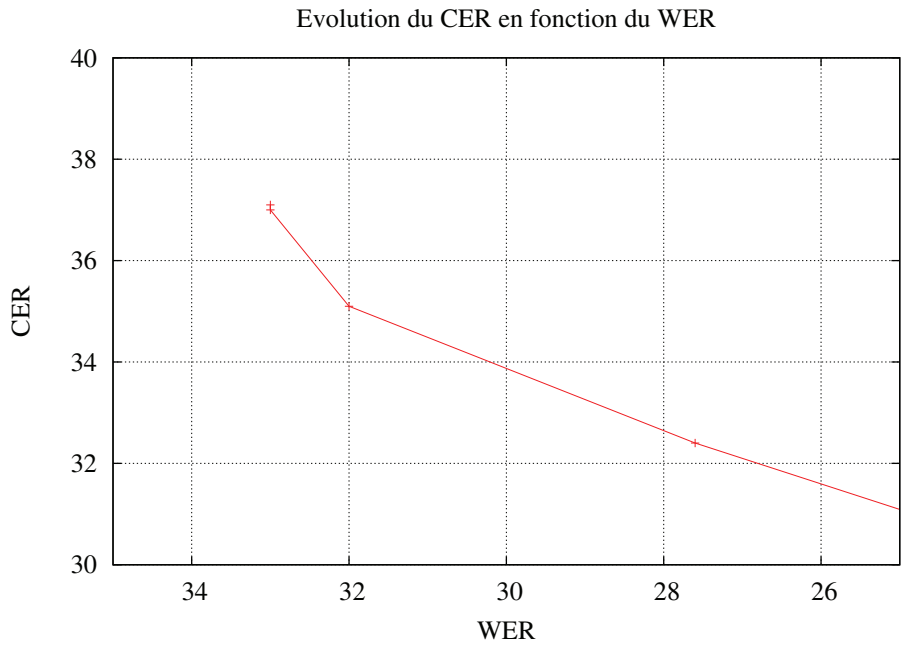

FIG. 4.3: *Courbe de l'évolution du CER en fonction du WER illustrant le tableau 4.4*

4.3.3 Synthèse sur les expériences issues de la RAP

Il y a un lien certain entre le résultat de reconnaissance conceptuelle et celui de la reconnaissance de la parole. L'amélioration de la transcription fournie par la combinaison LLC a un impact direct sur les interprétations sémantiques. Ce qui prête à penser que l'amélioration concerne des mots porteurs de sens.

Nous avons pu montrer à travers ces expériences que les *CER* et *WER* sont intimements liés. Il semble que la reconnaissance de la parole et le décodage conceptuel aient des performances linéairement dépendantes. En effet, lorsque la reconnaissance de la parole fonctionne mal, les défaillances se répercutent de manière uniforme sur les mots à reconnaitre et non sur certains mots en particulier. Inversement, les performances se retrouvent de manière uniforme sur les mots dans les performances du module de RAP. Les expériences montrent que ces différentes performances du modules de RAP sont répercutées sur le module de compréhension. D'après nos résultats, nous pouvons maintenant dire : meilleur est le *WER* de la reconnaissance de la parole, meilleur sera le *CER* du module de compréhension.

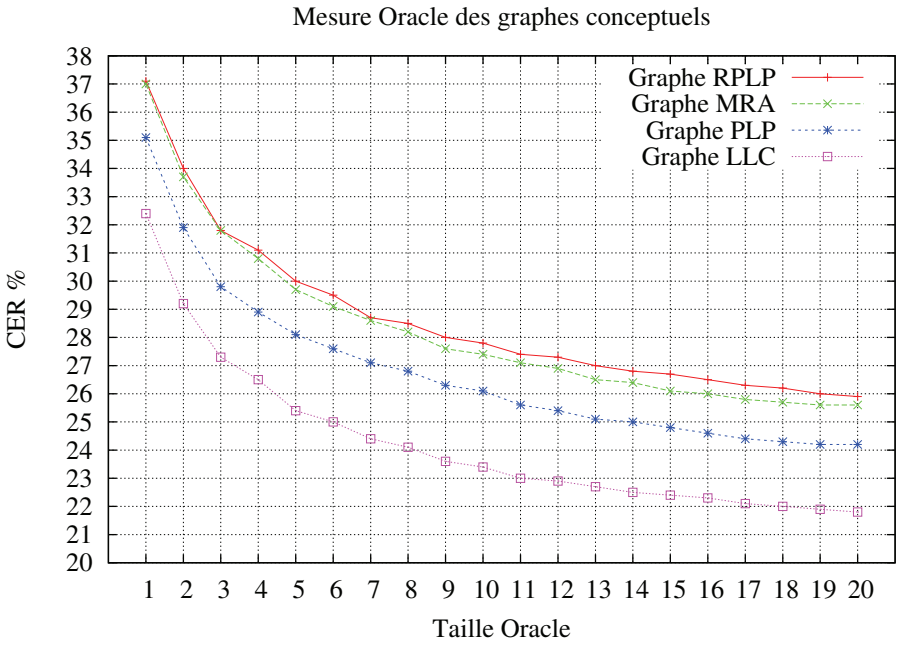

FIG. *4.4:* *Mesure Oracle des courbes RPLP, MRA, PLP et LLC*

4.4 Intégration de connaissances *a priori* dans un modèle stochastique

Le corpus d'apprentissage de 720 dialogues a été découpé en 7 parties, contenant respectivement 25, 50, 100, 200, 400, 600 et 720 dialogues. Ces différentes tailles permettent de mesurer l'impact sur les performances de la quantité de dialogues disponibles pour l'apprentissage des modèles. Cet apprentissage concerne deux types de modèles :

- les automates à états finis A_{c_i} représentant les concepts c_i, tels que présentés à la section 3.5 ;
- le modèle de langage présenté à la section 3.6 pour estimer la probabilité $P(t_{1,n}, w_{1,n})$.

Les performances sont mesurées par rapport au taux d'erreur sur les paires attribut/valeur. Un concept détecté est considéré comme correct uniquement si l'attribut du concept, ainsi que sa valeur normalisée, sont corrects d'après la référence. En alignant la chaîne de concepts détectés automatiquement et celle présente dans la référence établie manuellement, on calcule le Taux d'erreur Conceptuel (ou *CER*) suivant l'équation 2.20.

La campagne MEDIA s'étant d'abord focalisée sur le traitement de transcriptions manuelles, nous présentons ici les résultats obtenus sur ces transcriptions. Le graphe de mots G présenté à la section 3.6 est ainsi réduit à la seule chaîne de mots correspondants à la transcription manuelle de chaque énoncé. Le corpus de tests contient 200 dialogues représentant environ 3 000 énoncés et 8 500 occurrences de concepts.

Les expériences sont focalisées sur l'introduction de règles de grammaire manuelles parmi celles induites par un corpus d'apprentissage. Nous mesurons l'intérêt de cette méthode par rapport à la taille du corpus d'apprentissage. Les expériences seront menées sur la transcription manuelle du corpus de test. Le cadre de l'évaluation des expériences est le même que dans la section 4.2.

4.4.1 Performances *vs.* taille de l'apprentissage

Le tableau 4.5 présente le taux d'erreur sur les concepts (CER) en fonction de la taille du corpus d'apprentissage. La valeur du CER décroît de 46% jusqu'à 24% en augmentant la taille du corpus d'apprentissage, de 25 à 720 dialogues (soit environ 12K d'énoncés). Notons qu'un CER de 24% est comparable aux performances obtenues par les meilleurs systèmes testés lors de la campagne d'évaluation MEDIA.

Taille corpus	**25**	**50**	**100**	**200**	**400**	**600**	**720**
CER	46.0	43.7	34.6	31.9	26.3	24.2	24.3
correct	61.9%	63.3%	70.9%	74.7%	79.6%	80.8%	81.3%

TAB. 4.5: *Taux d'erreur sur les concepts (CER) du corpus de test et pourcentage de concepts corrects en fonction de la taille du corpus d'apprentissage (en nombre de dialogue)*

La décroissance du taux d'erreur ralentit fortement après 400 dialogues et se stabilise à 600 dialogues, indiquant qu'une augmentation de la taille du corpus d'apprentissage au-delà de 800 dialogues l'amélioration des performances n'est plus aussi significative. La taille optimale d'un corpus d'apprentissage pour l'approche présentée ici se situe ainsi autour de 400 dialogues. Le tableau 4.5 permet aussi d'estimer la taille minimale de corpus nécessaire afin d'obtenir un système présentant de bonnes performances. On remarque qu'un minimum de 100 dialogues est requis pour avoir un taux d'erreur inférieur à 35% et pour obtenir un pourcentage de concepts corrects d'environ 70%.

Il est intéressant de comparer cette courbe performance/taille de corpus avec celle présentée dans la description de la campagne MEDIA [Bonneau-Maynard *et al.*, 2005], mettant en relation le nombre de dialogues annotés manuellement et la mesure d'accord entre les annotateurs (*IAG*). Cette courbe montre qu'il a fallu environ trois itérations et une centaine de dialogues annotés pour obtenir un taux d'accord inter-annotateurs satisfaisant. Le manuel d'annotations n'a cessé d'évoluer et s'est stabilisé à la fin de l'annotation du premier lot du corpus MEDIA contenant 200 dialogues. On peut ainsi faire le parallèle entre la quantité de corpus nécessaire à la mise au point d'un manuel définissant les concepts d'une application donnée et celle nécessaire à l'apprentissage de

modèles statistiques permettant de modéliser ces mêmes concepts.

4.4.2 Ajout de connaissances *a priori*

Pour remédier au problème du manque de données, surtout dans les premières étapes du processus de développement d'une application où peu de données d'apprentissage sont disponibles, il est tentant d'ajouter des connaissances *a priori* dans les modèles de décodage conceptuel. L'approche présentée dans cette section permet d'effectuer directement cette intégration : les automates utilisés pour obtenir le graphe de concepts présentés dans la section 3.5 ne sont pas stochastiques. Ainsi, on peut ajouter directement au transducteur $T_{concept}$ d'autres automates représentant des connaissances *a priori* sur les concepts de l'application. Par exemple, certaines entités, comme les dates ou des entités numériques, se retrouvent dans de nombreuses applications et peuvent être facilement modélisées sous forme de grammaires régulières. Ces grammaires, une fois représentées sous forme d'automates, sont fusionnées avec celles dérivées du corpus d'apprentissage.

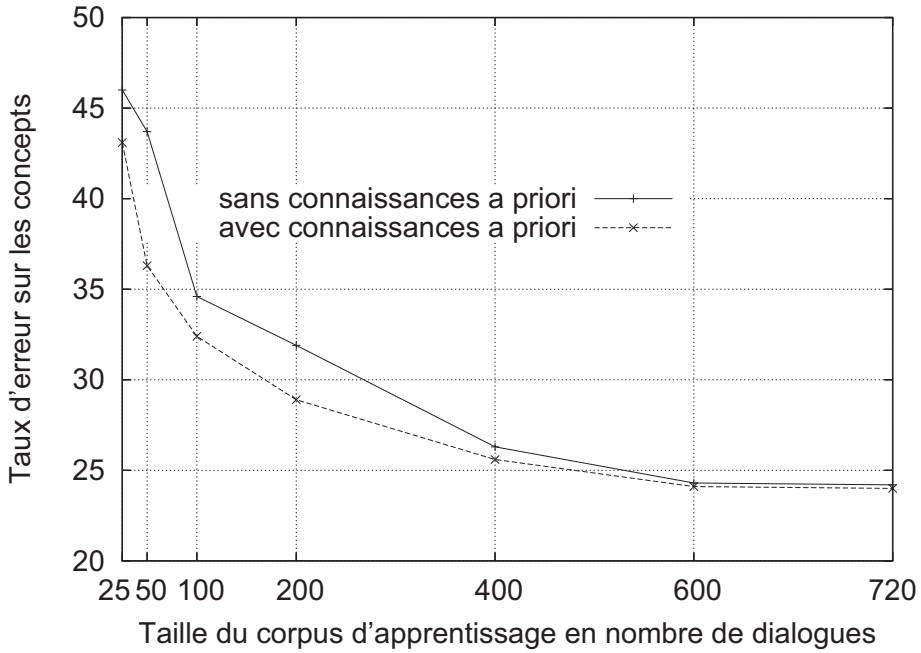

FIG. 4.5: *Taux d'erreur sur les concepts en fonction de la taille du corpus d'apprentissage, avec et sans connaissances a priori*

La figure 4.5 présente les courbes du CER en fonction de la taille du corpus d'apprentissage avec et sans connaissances *a priori*. Ces connaissances sont ici réduites à des

grammaires pour les dates et des entités numériques, telles que le nombre de chambres ou le nombre de personnes. Comme on peut le voir, ces grammaires permettent de pallier le manque de données dans les premières phases de l'apprentissage ; cet avantage s'estompe à mesure que la couverture du corpus d'apprentissage augmente. Par exemple, grâce à ces connaissances *a priori*, les performances du système avec un apprentissage sur 100 dialogues sont identiques à celles obtenues sans ces connaissances avec le double de dialogues d'apprentissage, réduisant ainsi de moitié l'effort d'annotation.

4.5 Intégration de l'historique dans les modèles de langage

L'approche intégrée de décodage conceptuel nécessite le traitement de graphes de mots issus du module de RAP. Ces graphes sont produits par le décodeur SPEERAL, toutes les opérations sur les graphes sont ensuite effectuées avec l'ensemble des outils de manipulation d'automates *AT&T FSM/GRM Library* [Mohri *et al.*, 2002]. Ces graphes nous permettent également de faire varier les probabilités des modèles. En effet, le but de cette méthode est d'étudier la corrélation entre l'historique du dialogue sur les hypothèses conceptuelles. Il serait intéressant d'adapter nos modèles de décodage conceptuel en fonction des différentes situations du dialogue.

4.5.1 Ètude de faisabilité

L'objectif de cette étude est d'améliorer le taux d'erreur conceptuel en fonction de l'historique de dialogue à partir du graphe de mot issu de la RAP. Nous avons décidé d'utiliser les entités nommées données par le Magicien d'Oz (ou *compère*) et annotées dans les transcriptions. Je suis parti du postulat que tout ce que le compère disait était forcément identifié et référencé par la machine. Les informations pertinentes fournies par le compère pour le dialogue et répétées par le locuteur sont les entités nommées relatives à des noms d'hôtel principalement. En effet, intuitivement, lorsque le locuteur veut effectuer une réservation, il ne connait pas le nom de l'hôtel. Il ne fait que répeter ce que le compère lui propose ; un exemple illustrant ce processus de dialogue est présenté dans l'encadré 4.6.

spk :	je voudrais réserver un hôtel à Avignon pour la mi-juillet
woz :	je vous propose l'hôtel Ibis Centre Gare à 45 euros la nuit et l'hôtel Mercure Pont d'Avignon à 160 euros la nuit
spk :	je prendrais alors le Centre Gare

TAB. 4.6: *Exemple de dialogue* `MEDIA` *incluant une répétition du nom d'hôtel.*

Dans un premier temps, il est intéressant de produire des statistiques concernant notre objectif. Nous pouvons observer le graphique 4.6 qui représente le nombre et la distance, en nombre de tours, entre la présence d'un nom d'hôtel répété par un locuteur

et la proposition du même nom d'hôtel par le compère dans le corpus d'apprentssage. Le graphique 4.7 présente ces mêmes statistiques pour le corpus de test.

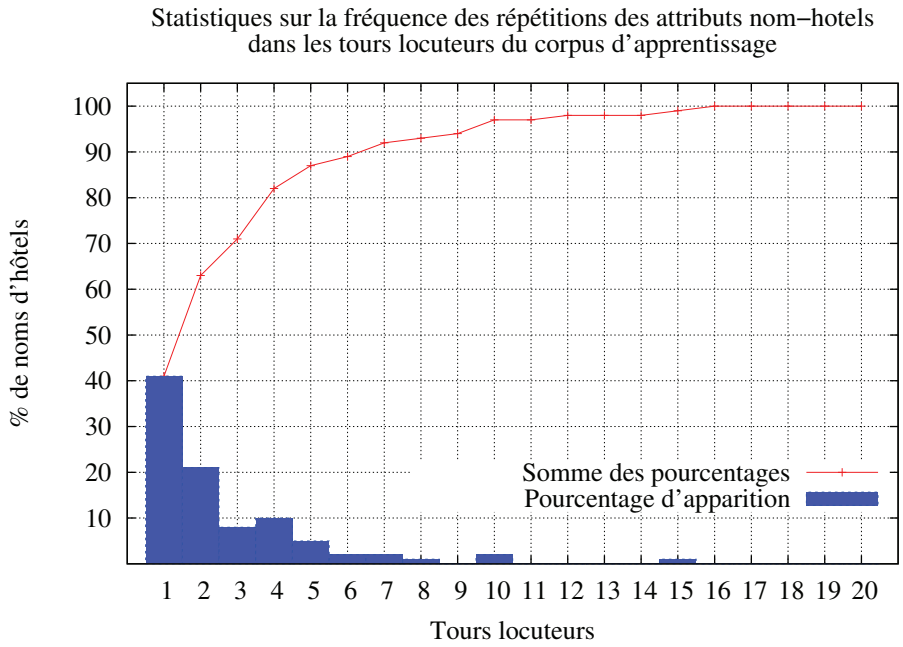

FIG. 4.6: *Pourcentage de la fréquence d'apparition des répétitions des attributs nom-hotel dans les tours locuteurs suivant un tour compère dans le corpus d'apprentissage*

Dans un premier temps, sur le corpus d'apprentissage, nous pouvons observer que plus de 40% des noms d'hôtels prononcés par le locuteur sont répétés par ce dernier dans le tour suivant la proposition du compère. Cette proportion augmente jusqu'à dépasser les 95% de noms d'hôtels au dixième tour locuteur (figure 4.6). Dans un second temps, nous observons dans le corpus de test que le même taux de répétition de nom d'hôtel, répété par le locuteur dans le tour suivant la proposition faite par le compère, est équivalent à celui du corpus d'apprentissage, tandis qu'à partir d'une distance de cinq tours, le taux de 95% de noms d'hôtel répétés se retrouve(figure 4.7).

Il nous faut effectuer une étude d'erreur sur les noms d'hôtel à partir des résultats du test du décodage conceptuel du graphe de mots issu de la RAP. Nous pouvons voir dans le tableau 4.7, les différents types d'erreurs par rapport au nombre d'occurrences des noms d'hôtel dans la référence.

Nous pouvons voir un fort nombre de substitutions qui représente près de 30% des noms d'hôtels à trouver, en comparaison du peu d'erreur d'insertion et du taux de 10% d'erreurs de délétion. Au total, nous avons un CER sur les noms d'hôtels qui s'élève à près de 46%.

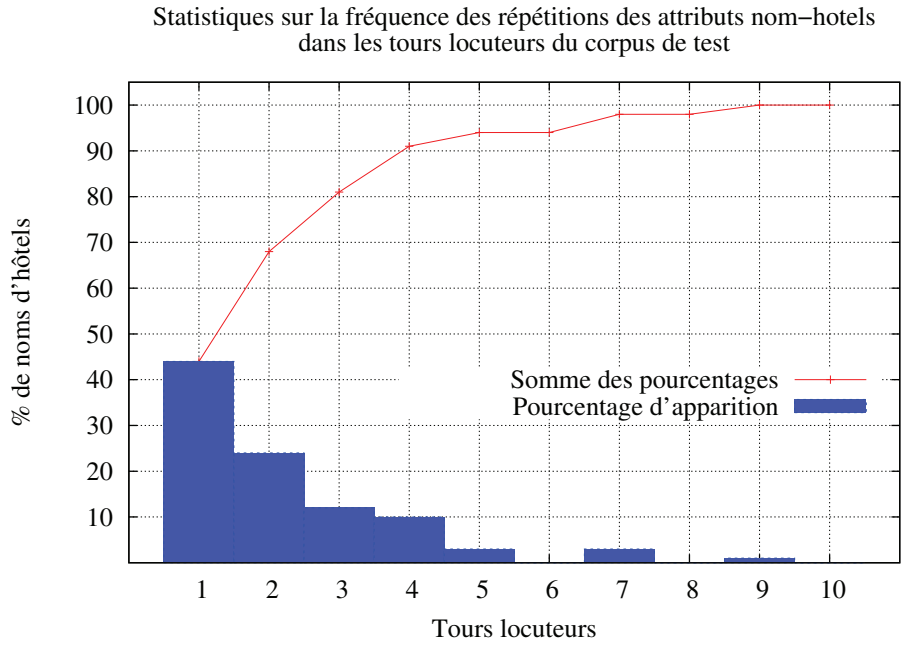

FIG. 4.7: *Pourcentage de la fréquence d'apparition des répétitions des attributs nom-hotel dans les tours locuteurs suivant un tour compère dans le corpus de test*

Type d'erreur :	Délétion	Insertion	Substitution	Total (CER)	Référence
Nbr d'erreur Int.	16	18	42	76	140
Nbr d'erreur Oracle Int.	12	7	33	52	140

TAB. 4.7: *Analyse du nombre des différents types d'erreurs sur les noms d'hôtels, suivant les différentes approches.*

Après une fine analyse de ces erreurs, il s'avère que parmi celles-ci, près de 18 noms d'hôtels, sur les 90 noms d'hôtels différents prononcés par le locuteur dans le corpus de test, ne sont pas présents dans le lexique du système de RAP. Il est évident que ces mots absents du lexique sont des erreurs que le système de compréhension ne pourra compenser. Ces mots hors-vocabulaire sont la source de 21 des erreurs sur les 140 occurrences de noms d'hôtels (plusieurs noms d'hôtels étant répétés plusieurs fois). Parmi ces substitutions, les autres noms d'hôtels restants sont présents dans les graphes de solutions du système de compréhension.

Nous sommes allés encore plus loin en produisant le tableau 4.8. Nous cherchons à savoir, lorsque nous rencontrons un nom d'hôtel dans notre treillis d'hypothèses (*Treillis*), si cet hôtel est présent dans notre meilleure hypothèse (*1-best*), s'il est présent dans la référence (*Référence*) et s'il a été prononcé par le compère dans le tour précédent

Recherche	**Présence**								
Tour compère précédent	1	1	1	1	1	/	/	/	∅
Tour locuteur (référence)	1	1	∅	∅	1	1	1	∅	1
Tour locuteur (1-Best)	1	∅	1	∅	∅	1	∅	1	∅
Tour locuteur (Treillis)	1	1	1	1	∅	/	/	/	∅
Nombre d'occurrence	26	12	3	3	21	82	58	18	140

TAB. 4.8: *Analyse des différents types d'erreurs sur les noms d'hôtels, suivant les différentes approches (1-Best et Treillis).*

ou non (première ligne du tableau). Nous pouvons constater, dans le tableau 4.8, les résultats d'analyses comparatives de corpus et de résultats. Un jeu de test correspond à une colone, les « 1 » symbolisent la présence du nom d'hôtel recherché dans l'ensemble correspondant (le nom de la ligne), et le « ∅ » l'absence de cet ensemble. Par exemple, il y a 9 noms d'hôtels présents dans le *treillis* d'hypothèses locuteur, dans la *Référence* et dans le tour compère précédent mais absent de la meilleure hypothèse locuteur (*1-Best*). Nous constatons que, sur les 140 noms d'hôtels, seuls 9 sont « oubliés » par la meilleure hypothèse d'interprétation de notre système, lorsqu'un nom d'hôtel est prononcé par le compère et que le locuteur le répète dans le tour suivant. Le système identifie dans son treillis 38 noms d'hôtels sur 59 prononcés par le compère et répétés par le locuteur. Soit plus de 64% des noms d'hôtels présents dans cette configuration dans le corpus de test. La possibilité de mettre une règle arbitraire suivante : « *Si un nom d'hôtel apparaît dans le treillis de mot et qu'il est présent dans le tour compère précédent, alors choisir le meilleur chemin passant par ce nom d'hôtel* » provoquerait 3 fausses acceptations et nous permettrait de retrouver 9 noms d'hôtels, dans le cas présent. Les gains possibles sont donc très intéressants en rapport des pertes possibles sur les noms d'hôtels. Malheureusement, les quantités prises en compte sont trop faibles pour pouvoir tirer des conclusions fiables.

4.5.2 Conclusions et perspectives de l'analyse d'erreur

Notre analyse d'erreur a montré qu'en conservant les 20 meilleures solutions du graphe (oracle de taille 20) nous couvrons plus de 67% des cas présentés. Une règle telle que nous l'avons exprimée permettrait de récupérer tous les noms d'hôtels présents dans les graphes et proposer les bonnes solutions. La contre-partie serait la généraion de 3 fausses reconnaissances. Ces résultats bien qu'encourageants ne permettent malheureusement pas de généraliser ce type de règle de manière fiable, car le nombre de cas présents dans corpus est trop faible.

Il serait intéressant de mettre en place des techniques utilisées dans la parole comme un modèle cache [Kuhn et De Mori, 1990] ou comme un *trigger*, utilisé récemment pour la traduction automatique [Lavecchia *et al.*, 2007], c'est-à-dire d'augmenter très fortement les probabilités d'apparition d'un nom d'hôtel prononcé par le compère lorsque celui-ci apparaît dans le treillis de solution. Une autre solution serait d'utiliser ces mêmes principes de modification de probabilités d'apparition d'un nom d'hôtel prononcé par le compère, lorsque, grâce aux actes de dialogues, on s'attend à avoir comme

réponse un nom d'hôtel et lorsque ce-dernier est présent dans le treillis de solution.

On peut s'interroger sur l'influence d'une telle proposition sur le treillis d'hypothèses. En effet, faire ressortir les noms d'hôtels de cette manière ne risque-t-il pas de générer d'autres erreurs sur les autres concepts ?

4.6 Conclusions et perspectives

Nous avons présenté dans les chapitres 3 et 4 un modèle de décodage conceptuel, basé sur une approche stochastique, utilisant des automates à états finis (ou *FSM*). Cette architecture nous permet une souplesse d'utilisation en utilisant aussi bien des séquences de mots, que des graphes de mots issus d'un processus de Reconnaissance Automatique de la Parole (RAP). L'un des principaux avantages de cette approche est de garder l'espace probabiliste des phrases produit en sortie du module de RAP et de le projeter vers un espace probabiliste de séquences de concepts. Ainsi, l'incertitude dans l'interprétation d'un message peut-elle être gardée plus longtemps pour être levée par des niveaux supérieurs d'interprétation intégrant le contexte du dialogue.

Les expériences menées sur le corpus `MEDIA` montrent que les performances atteintes par notre modèle sont au niveau des performances des systèmes ayant participé à la campagne d'évaluation dans les mêmes conditions d'évaluation que lors de la campagne `MEDIA`. En détaillant les performances de notre système en fonction de la taille du corpus d'apprentissage, on peut mesurer le nombre minimal ainsi que le nombre optimal de dialogues nécessaires à l'apprentissage des modèles. Il est particulièrement intéressant de mettre ces résultats en rapport avec ceux obtenus lors de la constitution du corpus `MEDIA`, notamment ceux sur le nombre de dialogues d'exemples nécessaires afin d'obtenir un manuel d'annotation stable. Les systèmes basés sur une modélisation explicite des connaissances ont besoin d'une telle spécification des concepts. Ce genre d'étude permet de relativiser l'argument présentant les méthodes à base d'apprentissage automatique comme nécessairement plus gourmandes en terme de corpus d'apprentissage.

Nous avons montré comment des connaissances *a priori* peuvent être intégrées dans nos modèles. Ces connaissances permettent d'augmenter significativement leur couverture et donc de diminuer de manière importante, à performance égale, l'effort de constitution et d'annotation du corpus d'apprentissage.

De plus, les performances du décodage conceptuel se dégradent linéairement en fonction du taux d'erreurs sur les mots. Nous avons cependant montré qu'une approche *intégrée*, cherchant conjointement la meilleure séquence de mots et de concepts, donnait de meilleurs résultats qu'une approche *séquentielle*. Il apparaît, en outre, que lors de la génération d'une liste de n-meilleures hypothèses, celles produites à partir du graphe de solutions de la méthode *intégrée* sont plus pertinentes que celles issues du graphe de solution de l'approche *séquentielle*.

Concernant l'utilisation de l'historique du dialogue, nous proposons de filtrer le

treillis d'hypothèses, d'intégrer un *trigger* ou un modèle cache dans le système de compréhension. Ces propositions sont à l'étude actuellement.

Chapitre 5

Routage d'appels et décodage conceptuel

Sommaire

5.1 Introduction

Le routage d'appel est une tâche qui consiste à labéliser ou à caractériser un message vocal. Chaque label correspond au service d'une application téléphonique. Ceci peut être considéré comme le premier niveau d'un processus de système de compréhension de la parole. Le type d'un appel correspond à une thématique générale d'un énoncé, par exemple : le type d'une requête dans un système de dialogue oral comme *How May*

I Help You ? [Gorin *et al.*, 1997]. Cette tâche a été bien étudiée et nombre de stratégies ont été proposées, la plupart assimilent le problème à une tâche de classification, utilisant des classifieurs dits « état de l'art » comme les Machines à Vecteur de Support appelés aussi *SVM* [Haffner *et al.*, 2003] ou des algorithmes de *Boosting* [Zitouni *et al.*, 2003].

Le principal point faible de ces techniques est la nécessité d'avoir un corpus annoté, sur lequel les méthodes de classification pourront être apprises.À cette fin, il faut collecter des exemples de requêtes d'utilisateurs associés à un type de requête, grâce, par exemple, à la méthode du Magicien d'Oz. Cette contrainte majeure vient du fait que la collecte de corpus est particulièrement coûteuse et quelquefois difficile à mettre en œuvre. Le processus d'annotations manuelles des données collectées est également très coûteux et est l'un des principaux freins au déploiement d'un système de dialogue type routage d'appel. Plusieurs études ont proposé de réduire cette dépendance aux annotations manuelles, voire de la supprimer, tout du moins pour la transcription des énoncés d'apprentissage. En effectuant, par exemple, une association directe entre les énoncés et le type, au niveau phonétique ou morphologique [Alshawi, 2003, Huang et Cox, 2003]. Cependant, toutes les études nécessitent un corpus d'apprentissage sur lequel sont apprises les associations entre les énoncés et les types d'appels.

Les travaux ici présentés vont plus loin dans la réduction de la collecte de données spécifiques et d'annotation afin de développer un nouveau système de classification d'appel. Nous présentons une méthode de développement d'un prototype sans collecte d'aucune sorte pour les corpus d'apprentissage du système de Reconnaissance Automatique de la Parole (RAP), d'une part, et, d'autre part, des modèles de classification. L'idée principale est de réutiliser des données d'entraînement existantes pour entraîner les modèles, sachant que l'on dispose de connaissances *a priori* sur la tâche à accomplir.

La partie expérimentale utilisée dans cette étude est un système de routage d'appels appliquée aux renseignements téléphoniques du Conseil Général de Vaucluse, qui a financé de cette thèse. En figure 5.1, nous avons un exemple de page HTML extraite de la FAQ, telle qu'on peut la voir sur le site.

Un système de routage d'appel est composé, dans notre cas, d'un module de reconnaissance automatique de la parole (RAP), d'un système de compréhension et d'un système de classification. Ces modules ont besoin des ressources pour pouvoir être paramétrés correctement.

Le module de RAP nécessite un modèle de langage, un modèle acoustique et un lexique. Nous ne possédons aucune ressource acoustique, nous ne pouvons donc pas adapter notre module de RAP avec un modèle acoustique, il nous faut en trouver un. Par contre, concernant le lexique et le modèle de langage, nous possédons des données à partir desquelles nous pouvons les extraire.

La compréhension nécessite un modèle de langage conceptuel et une ontologie. Au vu des données en présence il est possible de créer une ontologie et d'entraîner un modèle conceptuel sur ces données. Il nous faut néanmoins les annoter pour pouvoir entraîner notre modèle conceptuel.

Enfin, le système de classification nécessite un corpus d'exemple que nous ne pos-

sédons pas et qu'il faut créer à partir des données disponibles.

5.2 Récolte et traitement du corpus

5.2.1 Le corpus WEB

Au Quotidien (Famille/Santé/Travail)	Assurances	Élections
Environnement	Relations Extérieures	Impôts, Taxes & Douanes
Papiers & Administration	Personnes Handicapées	Retraite
Transports	Vacances-Loisirs	Vie Associative

TAB. *5.1:* *Thèmes utilisés comme classes dans le système de classification.*

Les seules données disponibles et associées au domaine sont celles du site web du Conseil Général de Vaucluse [1] et sa FAQ. Utilisant l'approche proposée dans le cadre du projet WebTalk [Feng *et al.*,], nous avons considéré le site web comme une structure de base de données d'information linguistique. Nous en avons besoin pour extraire les données nécessaires pour la construction d'un système de classification d'appels : un lexique, un modèle de langage pour la RAP et des jeux de *séquences de mots/concepts* pour le système de compréhension.

Le corpus, directement issu du traitement de la FAQ en ligne, contient tout le vocabulaire du domaine, les formulations des questions et leurs réponses associées. Nous en avons extrait un lexique de 2 872 mots. Voici un extrait de code HTML dans le tableau 5.2 à partir duquel nous effectuons une extraction de données.

Grâce à la structure de la FAQ, nous avons déterminé 12 classes pour notre système de classification et 44 concepts présentés respectivement dans le tableau 5.1 et 5.11.

Nous pouvons voir une vue d'ensemble du processus de création des données d'apprentissage dans la figure 5.2.

5.2.2 Le corpus EPAC

EPAC, ou *exploration de masse de documents audio pour l'extraction et le traitement de la parole conversationnelle*, est un projet sélectionné par l'ANR et associé au programme *Masse de Données - Connaissances Ambiantes* (*MDCA*). Ce projet associe quatre laboratoires : l'IRIT (Toulouse), le LI (Tour), le LIUM (le Mans) et le LIA. Ce projet a pour but de proposer des méthodes d'extraction d'informations et de structuration de données spécifiques aux documents audio.

Ce corpus a été sélectionné pour la richesse des expressions en langage naturel. Il nous permet d'introduire dans le corpus d'apprentissage la notion de langage naturel

[1] http ://www.vaucluse.fr

```
<a href=" ?xml=N358.xml&amp ;xsl=Noeud.xsl"> Carte nationale d'identité : personnes majeures (première demande et renouvellement)</a> </h3>
<ul>
<li><a href="#titrei__139141688_111"> Définition </a></li>
<li><a href="#titrei__139141688_172"> Bénéficiaire </a></li>
<li><a href="#titrei__139141688_202"> Durée de validité </a></li>
<li><a href="#titrei__139141688_226"> Coût </a></li>
<li><a href="#titrei__139141688_250"> Dépôt de la demande </a></li>
<li><a href="#titrei__139141688_311"> Pièces justificatives à fournir : </a></li>
<li><a href="#titrei__139141688_629"> Remise de la CNI </a></li>
<li><a href="#titrei__139141688_653"> Délais d'obtention </a></li>
</ul>
<a href="#document">Haut de page</a>
<div id="titrei__139141688_111"></div>
<h3> Définition </h3>
<p>La carte nationale d'identité (CNI) permet à son titulaire de certifier de son identité, même lorsqu'elle est périmée, sous réserve dans ce cas, que la photo soit ressemblante.</p>
<p>Elle n'est pas obligatoire.</p>
<p>En cours de validité, elle vaut document de circulation transfrontière pour se rendre dans les pays de l'Union Européenne et, sous certaines conditions, dans certains pays tiers (se rapprocher des services de l'ambassade ou du consulat du pays de destination).</p>
<p><img alt="" src="/include/Comarquage/images/l_excla.gif"><strong>Attention</strong> : le renouvellement et la première demande de CNI suivent la même procédure. Une nouvelle CNI est fabriquée pour son renouvellement.</p>
<p>Cependant, les pièces demandées peuvent varier selon les cas.</p>
```

TAB. 5.2: *Exemple de code HTML issu d'une page web (figure 5.1) utilisée pour l'extraction des données liées au domaine.*

Carte nationale d'identité : personnes majeures (première demande et renouvellement)
Définition
Bénéficiaire
Durée de validité
Coût
Dépôt de la demande
Pièces justificatives à fournir :
Remise de la CNI
Délais d'obtention
Haut de page
Définition
Attention

TAB. 5.3: *Exemple de ce que l'on extrait de la partie présentée dans l'exemple 5.2 et qui sera utilisé pour l'extraction des concepts.*

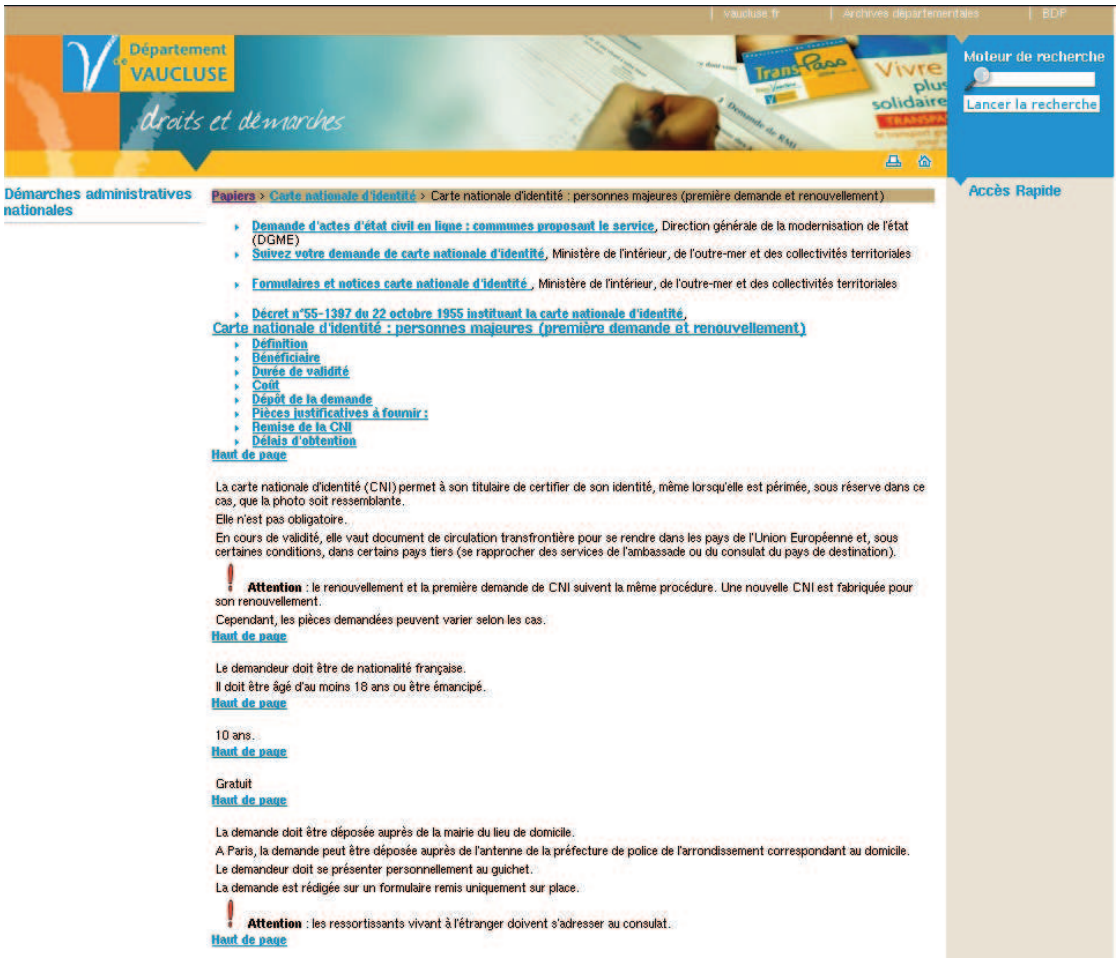

FIG. 5.1: *Extrait de la FAQ du site web.*

et de parole spontanée qui nous manque dans les autres corpus. Nous pouvons voir un extrait de ce corpus dans le tableau 5.4.

Dans le tableau 5.5, il est possible d'observer les séquences de mots correspondant au lexique. Ainsi, certains mots comme le mot *texte* ont un sens dans le contexte du corpus EPAC et qui dans l'application en auront un autre. Concernant notre exemple, dans le corpus EPAC, le mot *texte* fait référence au *texte* de la pièce de théâtre, alors que dans notre application le mot *texte* pourra être rattaché au domaine de la justice (i.e. *texte* de loi).

5.2.3 Le corpus RITEL

RITEL est un projet du LIMSI qui a pour but de réaliser un système de dialogue Homme-Machine, permettant d'interroger oralement une base de donnée, via un système de recherche d'information généraliste [Galibert *et al.*, 2005]. Le système pourra

- Didier Bezace , alors , dans votre théâtre de la Commune , vous mettez actuellement en scène deux pièces : - oui - "Chère Éléna Serguéfné... - oh ! Serguéiévna - Serguéiévna ! je savais que j'allais buter là-dessus ! Serguéiévna - et encore (il) y a le nom de l'auteur qui est pire - on en parlera tout à l'heure , justement . - c'est un texte contemporain euh russe - euh donc ce sera pour la deuxième partie de l'émission oui absolument euh spécial patinage pour moi hmm hmmmm - et puis seconde pièce que vous mettez en scène , "le square" de de Marguerite Duras - un texte finalement assez peu connu de cet auteur qui a marqué la fin du vingtième siècle oui oui - euh Duras pour vous euh c'est à la fois une première fois et et pas tout à fait - non , pas tout à fait oui . - pas tout à fait une première (en)fin le le le projet de monter "le square" est très ancien pour moi ; c'est un texte que qui m'a toujours tenu à cœur et qui m'a toujours euh paru important je dirais que c'est pour moi un texte du du de la littérature du vingtième siècle qui est très important et euh le projet a été remis pour un tas de raisons et puis voilà i(l)... sort maintenant comme ça arrive souvent alors effectivement euh j'ai j'ai adapté et mis en scène euh (il) y a je sais pas combien de temps une dizaine d'années ou peut-être un peu plus - en 92 - en 92 voilà bon euh voilà merci de me le rappeler euh... des conversations entre euh Marguerite Duras et... François Mitterrand qui avaient été publiées dans "l'Autre journal" et euh ce spectacle s'appelait "Marguerite et le président " euh il a eu beaucoup de succès et a été peu joué parce que il avait comme particularité que Marguerite était jouée par une petite fille de douze ans elle l'a vue , elle , Marguerite, ce spectacle , dont elle était très très amoureuse et euh et à partir de là euh j'ai su qu'un jour je ferai "le square" , voilà

TAB. 5.4: *Extrait du corpus EPAC : une conversation entre un journaliste et un directeur de théâtre.*

a	non
actuellement en scène deux pièces	oh
alors	on en
ce spectacle	ou
c'est un texte	oui
c'est un texte que qui	oui absolument euh spécial
dans	parce que il avait comme
dans votre théâtre de la	par une petite fille de douze ans elle
de de	pas tout à fait oui
de raisons et puis voilà	pas tout à fait une première
des	pièce que vous
dont elle était très très	pour moi hmm
elle	pour un
en	pour vous euh c'est à la fois une première fois et et pas tout à fait
en scène	que
entre euh	qui avaient été
est très ancien pour moi	qui est pire
était	qu'un jour je
et encore	sort maintenant comme ça arrive souvent alors effectivement euh
et euh ce spectacle	toujours euh paru important je dirais que c'est pour moi un texte du du de la littérature du vingtième siècle qui est très important et euh le projet a été
et euh et à partir de là euh	toujours tenu à cœur et qui
et le président	tout à l'heure
et mis en scène euh	un peu plus
et puis	un texte finalement assez peu connu de cet auteur qui a marqué la fin du vingtième siècle oui oui
euh	voilà
euh donc ce sera pour la deuxième partie de	voilà bon euh voilà merci de me le
euh il a eu beaucoup de succès et a été peu	vous
je	vue
justement	y a je sais pas combien de temps une
le le le projet de monter	y a le nom de

TAB. 5.5: *Texte conservé de l'extrait du corpus EPAC 5.4 pour la création du modèle de langage.*

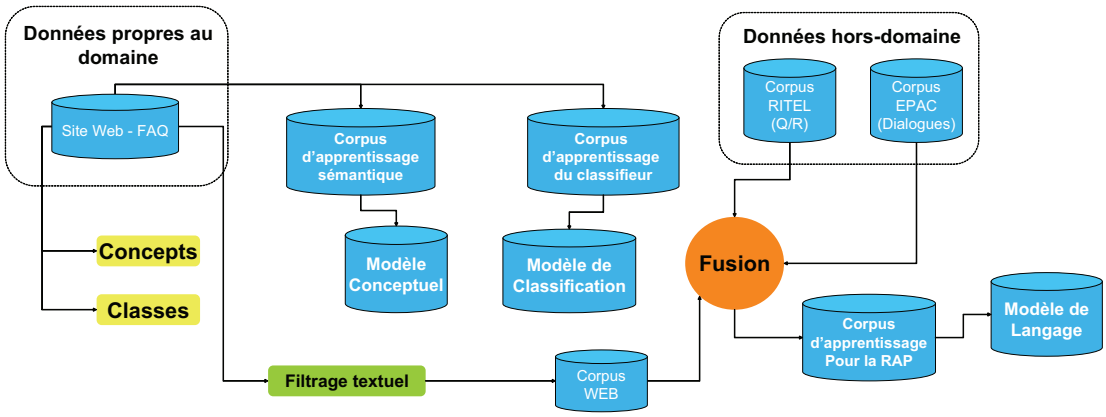

FIG. 5.2: *Vue d'ensemble du système d'extraction des données d'apprentissage du système.*

interagir avec l'utilisateur afin d'affiner ou de préciser la recherche (Extrait du corpus dans le tableau 5.6).

Le corpus issu de ce projet est composé de questions et de réponses. Les questions, une fois extraites du corpus, vont composer le nouveau corpus issu du corpus originel RITEL. Ce nouveau corpus contient les informations de formulation des questions. C'est cette notion de formulation de questions qui a retenu notre attention. En effet, notre système de routage d'appel est voué à utiliser les questions des utilisateurs pour caractériser les appels. Nous avons donc besoin d'informations linguistiques spécifiques aux questions, chose que ce corpus nous apporte.

5.2.4 Création du corpus d'apprentissage à partir des corpus à disposition

La création du corpus d'apprentissage est le résultat d'un mélange des trois corpus décrits précédemment. Ce mélange ne se fait pas au hasard, nous utilisons la mesure de la perplexité du modèle de langage issu du mélange de corpus. Afin d'aller au plus simple et au plus rapide, nous avons concaténé les trois corpus et avons extrait le modèle de langage pour le système de RAP. Nous pouvons voir, dans le tableau 5.8, les résultats de cette perplexité sur la transcription manuelle de notre corpus de test. Il est couramment admis qu'une perplexité en dessous de 100 points est considérée comme bonne.

5.2.5 Le corpus de test

Bien que nous n'utilisions pas seulement des données d'apprentissage spécifiques à la tâche, nous avons besoin de données de test afin d'évaluer notre système. Nous avons donc enregistré un corpus appelé *C_TEST*, contenant 216 messages prononcés par 10 locuteurs différents (hommes et femmes). La seule contrainte donnée aux locuteurs était d'exprimer une ou plusieurs requêtes correspondant à l'un des 12 thèmes.

- oui je voudrais connaître la date de construction du mur de Berlin
- je voudrais connaître la date de naissance du général de Gaulle
- je voudrais savoir où a été signée l' armistice de la première guerre mondiale
- quand a été assassiné J F K
- Julien Duvivier a -t-il réalisé les Don Camillo
- ah quelle est la hauteur de la tour Eiffel
- non je ne veux pas d' informations sur Benedetti je voudrais une information sur le dernier prix Nobel de la paix
- je m' intéresse à l' histoire et plus particulièrement à l' architecture et je je voudrais avoir qui a construit la chapelle Sixtine et à quelle époque
- je voudrais savoir quelle est la capitale des Canaries
- quel roi anglais a été décapité
- quand Paris a été libérée
- donc je c' est une question cinéma j' aimerais connaître le nom de l' acteur et le titre du film dans lequel on voit on voit l' image de d' un homme accroché à une aiguille de pendule
- j' aimerais savoir si l' Ethiopie est le pays le plus peuplé en Afrique
- j' aimerais savoir combien de langues officielles possède le Canada
- j' aimerais savoir le nombre d' enfants enfin combien Romy Schneider l' actrice a eu d' enfants
- sur un dieu qui a été condamné par Zeus à porter la voûte du ciel
- j' aimerais connaître le nom de l' artiste qui a créé le sacre du printemps
- non je voudrais connaître le titre du roman d' Alain Fournier
- non j' aimerais savoir quelle est la catastrophe naturelle qui fut la première à avoir été mise en poème
- j' aimerais connaître le nom de la capitale du Sri-Lanka
- c' est une question de littérature j' aimerais savoir si Trivelin est un personnage de Marivaux Molière ou Claudel
- j' aimerais savoir de quelle nationalité est l' écrivain Kessel
- j' aimerais connaître le nom du lac le plus profond du monde
- j' aimerais savoir en quelle année est mort Einstein
- j' aimerais savoir de quelle manière James Dean est mort
- de quel de quel pays Washington est la capitale
- où se trouve le quartier de Manhattan s'il-vous-plaît
- combien de films Hitchcock a -t-il tourné
- j' aimerais savoir quel est le pays qui a participé à tous les jeux olympiques d' hiver
- j' aimerais savoir quel est le laps de temps pendant lequel la tortue de mer couve ses oeufs

TAB. *5.6:* *Extrait du corpus RITEL : un ensemble de questions et de réponses sur aucun domaine précis.*

a	j' aimerais savoir en quelle année est mort
a été	j' aimerais savoir le nombre d' enfants enfin combien
a eu d' enfants	j' aimerais savoir quel est le
ah quelle est la hauteur de la tour	j' aimerais savoir quel est le pays qui a participé à tous les jeux olympiques d'
à porter la	j' aimerais savoir si l'
à une	je m' intéresse à l' histoire et plus particulièrement à l' architecture et je je voudrais avoir qui a
c' est une question de littérature j' aimerais savoir si	je voudrais connaître la date de naissance du général de
combien de films	je voudrais savoir où a été
de	je voudrais savoir quelle est la capitale des
de la	je voudrais une information sur le dernier prix
de la première guerre mondiale	l'
de mer	la chapelle
de quel de quel pays	la première à avoir été mise en
de temps pendant lequel la	le
donc je c' est une question cinéma j' aimerais connaître le nom de l' acteur et le titre du film dans lequel on voit on voit l' image de d' un homme	le plus
du	non j' aimerais savoir quelle est la catastrophe naturelle qui
du monde	non je ne veux pas d' informations sur
en	non je voudrais connaître le titre du roman d'
est la capitale	ou
est le pays le plus	oui je voudrais connaître la date de construction du mur de
est mort	où se trouve le quartier de
est un personnage de	quand
et à quelle époque	quand a été
j' aimerais connaître le nom de la capitale du	quel roi anglais a été
j' aimerais connaître le nom de l' artiste qui a créé le	qui a été condamné par
j' aimerais connaître le nom du	réalisé les
j' aimerais savoir combien de	ses
j' aimerais savoir de quelle manière	sur un
j' aimerais savoir de quelle nationalité est l' écrivain	tourné

TAB. 5.7: *Extrait du corpus RITEL 5.6 conservé pour la création du modèle de langage.*

Après une analyse *a posteriori* du corpus de test, nous avons découpé ce corpus en 3 niveaux de difficultés :

- le premier correspond aux messages de difficulté facile, ce sont des messages courts exprimant sans détours une requête précise. Par exemple : *J'ai une question à propos de ma retraitre*. Un jeu de 77 messages sont considérés comme faciles et rassemblés dans le sous-groupe C_TEST_facile.
- le second correspond aux messages de difficulté moyenne, ce sont des messages plus longs avec quelques incises, hors-sujet ou commentaires mais exprimant une requête précise. Par exemple : *Je viens juste d'acheter une nouvelle voiture et je voudrais savoir comment l'immatriculer*. Il y a 94 messages qui sont associés à ce niveau de difficulté et rassemblés dans le sous-groupe C_TEST_moyen.
- le dernier correspond aux messages difficiles, ce sont des messages très longs (parfois deux fois plus longs que le niveau de difficulté précédent) avec des incises, hors-sujet ou commentaires n'ayant que peu de rapport avec le sujet de la requête, cette dernière étant imprécise. Par exemple : *Je viens de d'emménager sur Avignon et j'avais demandé à une personne de vos service la dernière fois si...* Il y a 45 messages classés difficiles et rassemblés dans le sous-groupe $C\_TEST\_difficile$.

Toutes les évaluations seront faites à partir de ce corpus de test et des niveaux de difficulté associés.

ML	WEB	EPAC	RITEL	WEB+EPAC	WEB+RITEL	WEB+EPAC+RITEL
Nbr Mots	16k	114k	46k	130k	62k	176k
Perplexité	198	369	345	126	105	96

TAB. 5.8: *Perplexité mesurée sur le corpus C_TEST suivant le Modèle de langage utilisé.*

5.3 Création du modèle de langage pour la RAP

Le module de RAP utilisé pour ce travail de routage d'appel est SPEERAL. C'est le système de RAP du LIA. Faute de corpus audio, les modèles acoustiques utilisés sont ceux issus d'un apprentissage sur les corpus de la campagne ESTER. Ces modèles n'ont pas été adaptés à la tâche.

Les modèles de langage utilisés pour l'apprentissage sont issus d'un corpus d'apprentissage créé à partir de trois corpus : un corpus issu du site web, une partie du corpus RITEL et le corpus EPAC. Le premier est issu du traitement de la FAQ et contient les données du domaine concerné (*cf.* 5.2.1) ; le second est un corpus composé de transcriptions de paroles spontanées, issues de conversations en français entre des personnes (*c.f.* 5.2.2) ; le dernier est issu du projet RITEL [Rosset *et al.*, 2006], ce corpus nous permet d'introduire des transcriptions de paroles spontanées relatives aux formulations de questions (*c.f.* 5.2.3).

Les performances du système de RAP sur le corpus *C_TEST* sont présentées dans le tableau 5.9. Comme nous pouvons le voir, le taux de mots hors-vocabulaire (OOV) est de 6.7% et le taux d'erreur mots (WER) est de 54.2%. Il s'agit d'un taux d'erreur

C_TEST	*facile*	*moyen*	*diffile*	*tout*
taille	77	94	45	216
Nbr mots	808	1 209	914	2 931
Tx OOV	0.5%	10.8%	15%	6.7%
WER	28.1	52.7	73.5	54.2

TAB. 5.9: *Taux d'erreur mots obtenu sur les différents corpus de tests*

élevé, mais nous devons tenir compte du fait que nous ne disposons pas de corpus de développement pour adapter nos modèles de reconnaissance. Cependant, ce WER est dépendant de la complexité du message : pour les messages *faciles* le WER est aux alentours de 28.1% et il atteint 73.5% pour les messages *difficiles*, principalement à cause de leur contenu sans rapport avec la tâche et de leur taux élevé de mots hors-vocabulaire sur le corpus précis (15%).

Carte nationale d'identité : personnes majeures (première demande et renouvellement)

Définition
Bénéficiaire
Durée de validité
Coût
Dépôt de la demande
Pièces justificatives à fournir :
Remise de la CNI
Délais d'obtention

Haut de page

Définition
La carte nationale d'identité (CNI) permet à son titulaire de certifier de son identité, même lorsqu'elle est périmée, sous réserve dans ce cas, que la photo soit ressemblante.
Elle n'est pas obligatoire.
En cours de validité, elle vaut document de circulation transfrontière pour se rendre dans les pays de l'Union Européenne et, sous certaines conditions, dans certains pays tiers (se rapprocher des services de l'ambassade ou du consulat du pays de destination).
Attention : le renouvellement et la première demande de CNI suivent la même procédure. Une nouvelle CNI est fabriquée pour son renouvellement.
Cependant, les pièces demandées peuvent varier selon les cas.

TAB. 5.10: *Autre exemple d'extraction de données de la partie présentée dans l'exemple 5.2 et utilisée pour le modèle de langage du système de RAP.*

5.4 Création d'un étiquetage sémantique spécifique

Le routage d'appel est généralement considéré comme une tâche de classification. Ce processus de classification peut être directement appliqué aux mots sortis de la RAP ou en utilisant une étape intermédiaire : un système de compréhension qui traduira les séquences de mots en unités sémantiques appelés *concepts*. Cette approche est utilisée

dans le projet européen LUNA[2] où la première étape est de traduire des séquences de mots en concepts.

Afin d'obtenir les différents concepts pour notre tâche de classification, nous avons utilisé le processus suivant :

- nous avons utilisé la même méthode d'extraction de données que pour le corpus d'apprentissage du modèle de langage pour la RAP : nous avons utilisé les connaissances d'autres systèmes antérieurs afin de construire notre ontologie. Grâce à notre précédente contribution à la campagne TECHNOLANGUE/EVALDA/MEDIA (*cf.* chapitre 4) nous possédons une base de concepts génériques (tels que les dates, les nombres...), que nous avons réutilisés pour créer notre première liste de 24 concepts (*cf.* tableau 5.11).
- Ensuite, nous avons procédé à une opération d'extraction semi-automatique des mots compris entre les marqueurs *HTML* suivants : gras <B>, italique <I> et liens <A>. Grâce à l'utilisation d'une mesure de *TF.IDF* [Savoy, 2006] en combinaison avec une matrice de confusion, nous avons extrait 20 concepts spécifiques au domaine de la tâche, que nous avons ajoutés à notre ontologie générique. Nous pouvons voir un exemple du processus d'extraction de concepts dans l'encadré 5.3.

Concepts génériques			
localisation	adresse	arrondissement	code postal
département	pays	quartier	pays
rue	ville	nombre	pourcentage
nom	montant	réponse	requête
temps-année	temps-age	temps-date	temps-heure
temps-jour	temps mois		
Concepts Spécifiques			
Administrations	Armée	Assurance	Associations
Banques	Citoyenneté	Enseignement	Europe
Famille	Faune-Flore-Nature	Formation	Handicap
Impôt Taxe et Douane	Justice	Logement	Medical-Santé
Papiers	Travail	Detente	Véhicules

TAB. *5.11:* *Liste de concepts extraite de MEDIA et du corpus WEB*

Adaptation du module de compréhension

Les travaux effectués par AT&T sur *How May I Help You ?* [Gorin *et al.*, 1997] utilisent un module de compréhension pour leur système de routage d'appel. Nous disposons du système de compréhension créé et adapté pour la campagne MEDIA (chapitres 3 et 4), ainsi que les ressources nécessaires pour l'adapter à cette nouvelle tâche : un corpus annoté semi-automatiquement, des règles de complétions manuelles, un lexique et une ontologie. La mise au point nécessite une adaptation sur un corpus de développement, notamment, pour déterminer les bons coefficients de mélange entre le module de RAP et celui de la compréhension.

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

5.5 Classification thématique d'appels

La tâche de routage d'appels est, dans le cas présent, assimilée à une tâche de classification. Grâce à la FAQ du site web, nous disposons d'une collection de questions correspondant à une classe. Cette collection va nous servir de corpus d'apprentissage pour le système de classification. Nous appellerons le corpus d'apprentissage issu de la FAQ du site web, le corpus FAQ. Nous avons obtenu de cette manière 1 980 paires *questions/labels*. Il est à noter que nous avons deux corpus distincts : un corpus FAQ sans annotation sémantique pour l'apprentissage du classifieur sur les mots et une version annotée pour l'apprentissage du classifieur sur les mots et les concepts associés. Nous utilisons, pour nos expériences sur la classification, *BoosTexter*, un classifieur basé sur le texte et utilisant des algorithmes issus de AdaBoost [Schapire et Singer, 2000].

5.5.1 Classification sur les mots

Suivant le paradigme du *leave-one-out*, nous avons entraîné un modèle de classification sur les mots. Le principe du *leave-one-out* est le suivant : pour chaque exemple du corpus d'apprentissage, nous utilisons les 1 979 autres phrases comme corpus d'apprentissage. Ce paradigme nous permet de mesurer la cohérence de notre corpus. Le résultat de cette expérience a donné un nombre de phrases avec une classification correct de 1 601 sur 1 980 soit plus de 80% en utilisant les 12 classes du tableau 5.1. Malgré la petite taille du corpus d'apprentissage, nous avons considéré que le taux de classification correct était satisfaisant. Nous avons un extrait du corpus d'apprentissage dans l'encadré 5.12.

Phrase d'apprentissage	*Classe*
épargne à court terme livret jeune	Au_quotidien
quelle est la fiscalité de l' assurance vie lors d' un décès	Au_quotidien
logement assurance chômage	Assurances
garantie responsabilité civile	Assurances
réintégration des travailleurs ayant effectué le service national	Au_quotidien
récépissé de demande de titre de séjour	Relations_Extérieures
quelle est la durée de validité d' un extrait d' acte de naissance ou d' un autre acte d' état civil	Papiers_et_administration
dans quelles conditions un assuré pris en charge au titre de l' assurance maladie peut il prétendre à l' assurance invalidité	Personnes_handicapées
vol du permis de conduire établissement du duplicata	Transports
associations de défense de l' environnement et du cadre de vie	Vie_associative

TAB. 5.12: *Extrait du corpus d'apprentissage du classifieur : seuls les mots servent à caractériser une classe.*

5.5.2 Utilisation des concepts pour la classification

En utilisant l'étiquetage sémantique dans la classification, nous comptons améliorer le taux de classification des phrases prononcées. A cette fin, nous utilisons la mé-

Apprentissage mots-concepts	*Classe*
livret <Banques> jeune <Famille>	Au_quotidien
fiscalité <Impôts_taxe_et_douanes> assurance <Banques> vie <Banques> un <Nombre> décès <Famille>	Au_quotidien
logement <Logement> assurance <Assurances> chômage <Travail>	Assurances
garantie <Assurances> responsabilité <Justice> civile <Justice>	Assurances
travailleurs <Travail> service <Armée> national <Armée>	Au_quotidien
titre <Papiers> de <Papiers> séjour <Papiers>	Relations_Extérieures
un <Nombre> acte <Papiers> naissance <Famille> un <Nombre> acte <Papiers> état <Papiers> civil <Papiers>	Papiers_et_administration
un <Nombre> assurance <Médical-Santé> maladie <Médical-Santé> assurance <Assurances> invalidité <Handicap>	Personnes_handicapées
vol <Assurances> permis <Papiers> de <Papiers> conduire <Papiers>	Transports
associations <Associations> défense <Armée> environnement <Faune-Flore-Nature>	Vie_associative

TAB. *5.13:* *Extrait du corpus d'apprentissage du classifieur : les concepts et les mots supports du concept servent à caractériser une classe.*

thode d'analyse sémantique de la parole telle que nous l'avons décrite dans la section 3.2. Pour vérifier la cohérence du corpus annoté sémantiquement, nous avons suivi le même paradigme du *leave-one-out* et avons obtenu 1 714 phrases classées correctement sur 1 980, soit un taux de classification de 86% avec nos 12 classes. Cette amélioration brute de 6% par rapport à la classification avec les mots seuls est encourageante dans notre démarche. Les expériences de *leave-one-out* ont été faites sur de la transcription manuelle, dans nos expériences, nous utiliserons les sorties du module de RAP. Il est à noter que dans cette étude, le but est d'avoir le meilleur taux de classification, pas d'avoir le meilleur taux d'erreur conceptuel. Ainsi nous ne considérons dans nos expériences que le taux de classification. Nous avons un extrait du corpus d'apprentissage avec les annotations sémantiques dans l'encadré 5.13.

Pour nos expériences, sur le corpus C_TEST, nous allons comparer trois approches :

- la *base* de l'expérience : dans cette approche nous effectuons notre classification sur la meilleure hypothèse du module de RAP, avec seulement les mots reconnus ;
- *séquence de mots/concepts* : comme présenté dans la section 3.2, nous utilisons l'approche de décodage conceptuel de parole *intégrée* ; nous effectuons l'opération de classification sur la meilleure séquence de mots et de concepts produite par le décodage conceptuel sur le graphe de mots issus du module de RAP ;
- l'approche *sac de mots/concepts* : le taux d'erreur mot n'étant pas élevé, nous utilisons une nouvelle approche, un sac de concepts reconnus avec leurs mots-supports issus du graphe de mots.

Les deux dernières approches utilisent une méthode de classification sur les mots et sur les concepts en même temps. L'approche du *sac de mots/concepts* est décrite dans la partie suivante.

5.6 Utilisation d'un *sac de mots/concepts* pour la classification

Dans notre stratégie de décodage conceptuel, notre interprétation est assimilable à un processus de traduction en symboles sémantiques qui utilise des machines à état finis (*FSM*). Ces symboles sémantiques sont appelés des *marqueurs conceptuels* et sont notés C_i. Ils correspondent à chacun des 44 concepts définis dans l'ontologie (*cf.* partie 5.4). Chaque concept C_i est associé à une séquence de mots. Il existe un FSM pour chaque concept élémentaire. Chaque FSM est une approximation à états finis d'une grammaire en langage naturel. Ces FSM sont des *transducteurs* qui prennent en entrée des mots ou une séquence de mots et produisent en sortie un concept. Nous faisons l'union de tous les transducteurs en un seul que nous appellerons *FSM Concept*.

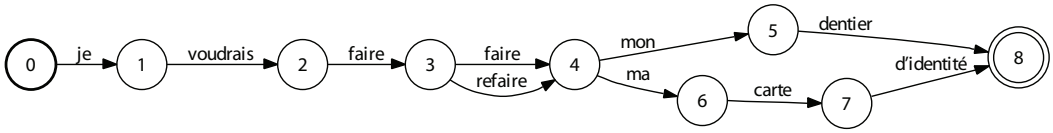

FIG. 5.3: *Graphe de mots issu du système de RAP pour la phrase prononcée « Je voudrais faire refaire ma carte d'identité ».*

Comme expliqué pour la méthode intégrée dans la partie 3.2, nous décodons un message d'un graphe de mots G_W, issu de la RAP. On peut voir un exemple de graphe de mots résultant de la reconnaissance de la phrase « Je voudrais faire refaire ma carte d'identité » avec la figure 5.3. Nous composons G_W avec le transducteur de concepts *FSM Concept*. Le résultat de cette composition est le transducteur T_{WC}. Chaque chemin de ce transducteur correspond aussi bien à une séquence de mots, si on considère les symboles d'entrée, qu'à une séquence de concepts, si on considère les symboles de sorties. Nous pouvons voir dans la figure 5.4 notre exemple annoté sémantiquement.

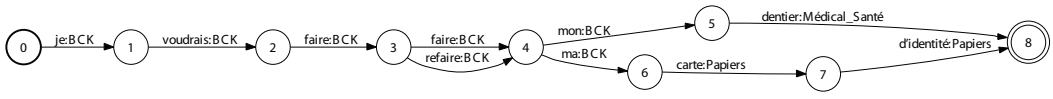

FIG. 5.4: *Graphe de mots/concepts issu du système de RAP et composé avec le transducteur de concepts pour la phrase prononcée « je voudrais faire refaire ma carte d'identité».*

C'est à partir de ce transducteur que nous extrayons les approches *séquence de mots/-concepts* et *sac de mots/concepts*, présentées précédemment. La première approche est classique dans notre recherche : il suffit d'extraire le meilleur chemin de T_{WC}. La seconde nécessite une méthode de filtrage que nous détaillerons dans la partie suivante.

5.6.1 Méthode de filtrage

L'approche du *sac de mots/concept* consiste à filtrer le transducteur T_{WC} avec n filtres correspondants aux n concepts de notre ontologie (ici 44 concepts). Chaque filtre F_i est un FSM simple : un accepteur qui accepte le ou les chemins correspondant au concept C_i. Nous pouvons observer l'exemple simple d'un filtre pour le concept « Papier » dans la figure 5.5. Le meilleur chemin de l'intersection entre T_{WC} et F_i est conservé. La probabilité $P(C_i|A)$ est obtenue avec la formule suivante :

$$P(C_i|A) = \frac{P(\hat{WC}_i|A)}{P(\hat{W}|A)} \tag{5.1}$$

Avec $\hat{W}$ étant la meilleure hypothèse de mot du graphe de mots G_W et A étant le signal de parole. Les probabilités $P(\hat{WC}_i|A)$ et $P(\hat{W}|A)$ sont données par le système de RAP et par les modèles de langage.

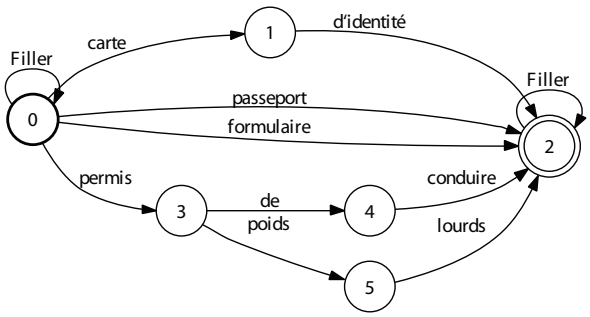

FIG. 5.5: *Exemple de filtre pour le concept « Papier».*

Ce filtrage du transducteur nous donne 44 probabilités $P(C_i|A)$, soit une pour chaque C_i, associée à la meilleure séquence de mots trouvée dans le graphe de mots G_W supportant le concept. Lors du processus de filtrage, si l'intersection d'un filtre F_k, associé à un concept k, avec T_{WC} est vide, alors la probabilité $P(C_k|A)$ associée à l'opération de composition est nulle.

Dans notre exemple (figure 5.5), le résultat du filtrage est le *sac de mots/concepts* suivant :
concept : Médical-Santé, dentier
concept : Papiers, carte, d'identité

Ce sac est filtré suivant la probabilité de $P(C_i|A)$ correspondante. Par exemple, si la probabilité $P(C_{Medical-Santé}|A)$ associée au concept *Médical-Santé* dans le transducteur T_{WC} est trop faible, alors le concept *Médical-Santé* sera écarté de notre solution filtrée.

5.7 Expériences

Nous avons expérimenté les trois approches présentées précédemment sur le corpus *C_TEST* et les sous-corpus *C_TEST_facile*, *C_TEST_moyen* et *C_TEST_difficile*. Le tableau 5.14 montre les résultats des expériences sur la classification d'appels. La colonne *Référence* concerne les expériences faites sur la transcription et la colonne *Reconnaissance Automatique de la Parole* est associée aux données issues du module de RAP. La sous-colonne *Mots* caractérise l'utilisation, comme support pour le système de classification, des mots seuls ; celle *Mots/concepts* aura pour support de classification la meilleure séquence de mots et de concepts ; enfin, l'étiquette *Sac de Mots/concepts* concerne l'utilisation de l'approche de *Sac de mots/concepts* décrite précédemment. Cette dernière approche ne sera appliquée que sur le treillis de mots issu du module de RAP.

Corpus	*Référence*		*Reconnaissance Automatique de la Parole*		
	Mots	*Mots/concepts*	*Mots*	*Mots/concepts*	*Sac de Mots/concepts*
C_TEST_facile	79.2	79.2	70.1	74.0	**80.5**
C_TEST_moyen	47.5	80.8	65.9	69.1	**71.2**
C_TEST_difficile	51.1	73.3	40.0	55.5	**57.7**
C_TEST	71.3	78.7	62.1	68.1	**71.7**

TAB. 5.14: *Résultat de la classification suivant la difficulté du corpus et l'introduction d'information sémantique.*

Comme nous pouvons le constater, l'ajout de l'information sémantique aux mots permet d'améliorer les résultats de classification par rapport à l'utilisation de mots seuls. L'approche de *Sac de Mots/concepts* rend le processus de classification plus robuste et moins dépendant d'une mauvaise transcription, produite par des modèles de RAP insuffisamment entrainés.

5.8 Conclusion

Ce chapitre a montré l'intérêt d'intégrer une tâche de classification d'appel dans un processus de compréhension de la parole spontanée. Nous avons appliqué les connaissances et les approches développées lors de la campagne `MEDIA` à une tâche de classification. La classification conceptuelle proposée dispose de très peu de données d'apprentissage. De plus, elle est intégrée au processus de reconnaissance de la parole. Cette approche nous permet de conserver l'espace de recherche probabiliste sur les séquences de mots produites par le système de RAP. Cet espace est projeté vers un espace de recherche probabiliste sur les séquences de concepts. Ces dernières, une fois filtrées, peuvent être exploitées en utilisant un classifieur grâce à la stratégie de *sac de mots/concepts*. Malgré de fortes valeurs de WER, les taux de classification d'appels obtenus sur le corpus de test sont, avec cette méthode, encourageants.

Chapitre 6

Conclusion

Après avoir exposé les difficultés de la langue naturelle orale et les différents principes qui y sont rattachés, nous avons proposé un système de compréhension de la parole. Ce système proposé est issu d'une approche intégrant le module de reconnaissance automatique de la parole (RAP) et le module de compréhension. Grâce à cette approche, nous conservons l'espace de recherche probabiliste de mots généré par le module de reconnaissance que nous interprétons grâce au module de compréhension. Nous avons en sortie du module de compréhension un nouvel espace de recherche probabiliste conceptuel qui a conservé les informations relatives aux mots. Cette approche nous permet de garder l'espace probabiliste des phrases produit en sortie du module de RAP et de le projeter vers un espace probabiliste de séquences de concepts. Ainsi, l'incertitude dans l'interprétation d'un message peut-elle être gardée plus longtemps pour être levée par des niveaux supérieurs d'interprétation intégrant le contexte du dialogue.

Dans le soucis de valider notre approche, nous avons mené des expériences sur le corpus MEDIA dans les mêmes conditions d'évaluation que lors de la campagne MEDIA. Les résultats montrent que les performances atteintes par notre modèle sont au niveau des performances des systèmes ayant participé à la campagne d'évaluation.

On présente les méthodes à base de d'apprentissage automatique comme nécessairement plus gourmandes en terme de corpus d'apprentissage. En modifiant la taille du corpus d'apprentissage, on peut mesurer le nombre minimal ainsi que le nombre optimal de dialogues nécessaires à l'apprentissage des modèles. Ayant mis ces résultats en rapport avec ceux obtenus lors de la constitution du corpus MEDIA sur le nombre de dialogues d'exemples nécessaires afin d'obtenir un manuel d'annotation stable. Nous avons montré que nous pouvions diminuer la taille du corpus d'apprentissage en intégrant à notre modèle des règles simples, telles que la grammaire des nombres, les manières simples de présenter un noms d'hôtels, les noms de villes, etc. Ces expériences permettent de relativiser l'argument « Les méthodes à base de d'apprentissage automatique sont nécessairement plus gourmandes en terme de corpus d'apprentissage », lorsqu'on peut y intégrer une modélisation explicite de connaissances générales.

En utilisant différentes sorties du module de RAP, nous avons montré que les performances du décodage conceptuel se dégradent linéairement en fonction du taux d'erreurs sur les mots (WER). Nous avons cependant montré qu'une approche *intégrée*, cherchant conjointement la meilleure séquence de mots et de concepts, donnait de meilleurs résultats qu'une approche *séquentielle*.

Nous avons appliqué les connaissances et les approches développées lors de la campagne `MEDIA` à une tâche dialogue simple : le routage (ou classification) d'appels. Nous avons montré l'intérêt d'intégrer une tâche de classification d'appel dans un processus de compréhension de la parole spontanée. La seconde partie intéressante vient du fait que nous disposons de très peu de données d'apprentissage relatives au domaine de la tâche. En utilisant l'approche intégrée de décodage conceptuel, conjointement à un processus de filtrage, nous avons proposé une approche sous forme de *sac de mots*. Cette approche exploitée par un classifieur permet d'obtenir des taux de classification d'appels encourageants sur le corpus de test, alors que le WER est assez élevé.

6.1 Perspectives

Cette thèse a permis de mettre en relief l'intérêt d'intégrer le processus de Reconnaissance Automatique de la Parole (RAP) au module de compréhension de la parole. Le fait d'effectuer un décodage conceptuel sur un graphe de mots, tel que nous l'avons proposé, permet de conserver un espace de recherche complet sur les mots, plutôt que de conserver la meilleure hypothèse. Cette méthode permet, en partie, de récupérer certaines erreurs faites par le module de RAP. Il est évident que cette approche ne permettra pas de récupérer les erreurs engendrées par la présence de mots hors-vocabulaire. Avec des mesures de confiance sémantique, il serait intéressant de voir si nous pouvons localiser les passages comportant des mots hors-vocabulaire et le cas échéant tenter de relancer le module de RAP en y ajoutant des mots comme présenté dans [Oger *et al.*, 2008] ou de manière dynamique comme dans [Chung *et al.*, 2004].

Dans le cadre d'un système de dialogue complet, il serait intéressant d'exploiter l'historique du dialogue, afin de proposer au module de gestion du dialogue des informations complémentaires, lui permettant d'affiner ses décisions. En effet, l'analyse d'erreurs a montré l'intérêt de modifier les probabilités des treillis de mots avec des *triggers*, un modèle cache ou d'utiliser des règles arbitraires obligeant le passage dans une partie du graphe et se déclenchant sur la présence d'éléments déclencheurs (mots ou concepts) en fonction de l'historique. Ces techniques sont actuellement à l'étude.

Une partie du travail sur la classification d'appel a été de réutiliser des corpus d'apprentissages, permettant de limiter les coûts de collecte de corpus spécifique pour l'application. Généraliser ce principe permet de réduire le cout de conception et le temps de mise au point d'une application. Cette méthode de protoypage de rapide systèmes permet d'établir une base de fonctionnement qui sera affinée en intégrant de nouvelles données d'apprentissage.

Il reste le problème la collecte, la création et l'annotation de nouvelles données

d'apprentissage qui ont un coût. Ainsi une des possible amélioration serait une intégration semi-supervisée de ces nouvelle données d'apprentissage. À la manière de [Haghighi et Klein, 2006], cette méthode se fonde sur peu d'exemples et étend l'étiquettage à un corpus par des statistiques globales. Par rapport à un corpus d'apprentissage de taille plus vaste, cet article montre que les performances ne sont pas au rendez-vous. L'intérêt de cette méthode est qu'un annotateur peut observer et se servir du résultat de cette projection pour relancer l'apprentissage en y ajoutant des exemples plus représentatifs. De cette manière, le travail de l'annotateur semble plus rapide et plus facile.

Liste des illustrations

Liste des tableaux

Bibliographie

[Allauzen *et al.*, 2004] Cyril ALLAUZEN, Mehryar MOHRI et Brian ROARK (2004). « A general weighted grammar library ». *Proceedings od the Ninth International Conference on Implementation ant Application Automata (CIAA'2004)*. 48, 52

[Allen, 1988] James ALLEN (1988). *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc. 25

[Alshawi, 2003] H. ALSHAWI (2003). « Effective utterance classification with unsupervised phonotactic models ». *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 1–7. 78

[Antoine *et al.*, 2003] Jean-Yves ANTOINE, Jerome GOULIAN et Jeanne VILLANEAU (2003). « Quand le TAL robuste s'attaque au langage parlé : analyse incrementale pour la compréhension de la parole spontanée ». *TALN'2003*, pages 25–34, Batz-sur-Mer, France. 43

[Baker *et al.*, 1998] Collin F. BAKER, Charles J. FILLMORE et John B. LOWE (1998). « The berkeley framenet project ». *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90, Morristown, NJ, USA. Association for Computational Linguistics. 3, 22

[Bangalore et Johnston, 2004] Srinivas BANGALORE et Michael JOHNSTON (2004). « Balancing data-driven and rule-based approaches in the context of a multimodal conversational system ». Daniel Marcu SUSAN DUMAIS et Salim ROUKOS, éditeurs : *HLT-NAACL 2004 : Main Proceedings*, pages 33–40, Boston, Massachusetts, USA. Association for Computational Linguistics. 3, 30

[Barrault, 2008] Loïc BARRAULT (2008). *Diagnostic pour la combinaison de systèmes de reconnaissance automatique de la parole*. Thèse de doctorat, Université d'Avignon. 58

[Barrault *et al.*, 2008] Loïc BARRAULT, Christophe SERVAN, Driss MATROUF, Georges LINARÈS et Renato De MORI (2008). « Frame-based acoustic feature integration for speech understanding ». *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'08)*, Las Vegas. 58

[Blanche-Benveniste, 1997] Claire BLANCHE-BENVENISTE (1997). *Approche de la langue parlée en français*. Collection l'essentiel français édition. 20

[Bonneau-Maynard *et al.*, 2006] Hélene BONNEAU-MAYNARD, Christelle AYACHE, Frédéric BÉCHET, Alexandre DENIS, Anne KHUN, Fabrice LEFEVRE, Djamel MOSTEFA, Mathieu QUINARD, Sophie ROSSET, Christophe SERVAN et Jeanne VILLANEAU

(2006). « Results of french evalda-media evaluation campaign for litteral understanding ». *Language Resources and Evaluation Conference (LREC'06)*, pages 2054–2059, Gènes (Italie). 38, 62

[Bonneau-Maynard et Lefevre, 2005] H. BONNEAU-MAYNARD et F. LEFEVRE (2005). « A 2+1-level stochastic understanding model ». *ASRU*, Porto Rico. 37, 43

[Bonneau-Maynard *et al.*, 2005] Helene BONNEAU-MAYNARD, Sophie ROSSET, Christelle AYACHE, Anne KUHN et Djamel MOSTEFA (2005). « Semantic annotation of the french media dialog corpus ». *EUROSPEECH*, Lisboa, Portugal. 32, 35, 69

[Bousquet-Vernhettes, 2002] Caroline BOUSQUET-VERNHETTES (2002). *Compréhension robuste de la parole spontanée dans le dialogue oral homme-machine - Décodage conceptuel stochastique*. Thèse de doctorat, Université de Toulouse III. 19

[Cettolo *et al.*, 1998] Mauro CETTOLO, Roberto GRETTER et Renato DE MORI (1998). « Search and generation of word hypotheses ». Renato DE MORI, éditeur : *Spoken Dialogues with Computers*, chapitre 9, pages 257–309. Academic Press. 16

[Chappelier *et al.*, 1999] J.C. CHAPPELIER, M. RAJMAN, R. ARAGUES et A. ROZENKNOP (1999). « Lattice parsing for speech recognition ». *Proceedings of the 6th conference on Traitement Automatique du Langage Naturel TALN'99, Cargese, Corsica, France*. 44

[Charniak *et al.*, 1993] Eugene CHARNIAK, Curtis HENDRICKSON, Neil JACOBSON et Mike PERKOWITZ (1993). « Equations for part-of-speech tagging ». *11th National Conference on Artificial Intelligence*, pages 784–789. 47

[Chen et Goodman, 1996] Stanley F. CHEN et Joshua GOODMAN (1996). « An empirical study of smoothing techniques for language modeling ». A. JOSHI et M. PALMER, éditeurs : *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, USA. Morgan Kaufmann Publishers. 13

[Chung *et al.*, 2004] Grace CHUNG, Stephanie SENEFF, Chao WANG et Lee HETHERINGTON (2004). « A dynamic vocabulary spoken dialogue interface ». *ICSLP*, Jeju Island (Corée). 96

[Cohen, 1960] Jacob COHEN (1960). « A coefficient of agreement for nominal scales ». *Educationnal and Psychological Measurement*, pages 20 :37–46. 35

[Crabbe *et al.*, 2003] B. CRABBE, B. GAIFFE et A. ROUSSANALY. (2003). « Une plateforme de conception et d'exploitation de grammaire d'arbres adjoints lexicalises ». *TALN*. 36

[De Mori, 1998] Renato DE MORI (1998). *Spoken Dialogues with computers*. 1, 9, 17, 24, 26, 27, 30

[Denis *et al.*, 2006] Alexandre DENIS, Matthieu QUINARD et Guillaume PITEL (2006). « A deep-parsing approach to natural language understanding in dialogue system : Results of a corpus-based evaluation ». 36, 43

[Feng *et al.*,] J. FENG, S. BANGALORE et M. RAHIM. « WebTalk : mining Websites for automatically building dialog systems ». *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 168–173. 79

[Fillmore, 1968] Charles J. FILLMORE (1968). « The case for case ». Emmon BACH et Robert T. HARMS, éditeurs : *Universals of Linguistic Theory*, pages 1–88. Holt, Rinehart, and Winston, New York. 22

[Galibert *et al.*, 2005] Olivier GALIBERT, Gabreil ILLOUZ et Sophie ROSSET (2005). « Ritel : An open-domain, human-computer dialog system ». *INTERSPEECH*. 81

[Gemello *et al.*, 2006] Roberto GEMELLO, Franco MANA, Dario ALBESANO et Renato MORI, de (2006). « Multiple resolution analysis for robust automatic speech recognition ». *Computer Speech & Language*, 20(1):2–21. 59

[Gorin *et al.*, 1997] Allen L. GORIN, Giuseppe RICCARDI et Jeremy H. WRIGHT (1997). « How may i help you ? ». *Speech Communication*, 23(1-2):113–127. 4, 28, 78, 89

[Gupta *et al.*, 2006] N. GUPTA, G. TUR, D. HAKKANI-TUR, S. BANGALORE, G. RICCARDI et M. GILBERT (2006). « The AT&T spoken language understanding system ». *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, 14(1):213–222. 28

[Haffner *et al.*, 2003] Patrick HAFFNER, Gokhan TUR et Jerry WRIGHT (2003). « Optimizing SVMs for complex call classification ». *Proc. IEEE ICASSP'03*, Hong-Kong. 78

[Haghighi et Klein, 2006] Aria HAGHIGHI et Dan KLEIN (2006). « Prototype-driven learning for sequence models ». *NAACL06*. 97

[Haton *et al.*, 2006] Jean-Paul HATON, Christophe CERISARA, Dominique FOHR, Yves LAPRIE et Kamel SMAÏLI (2006). *Reconnaissance automatique de la parole*. Dunod. 10

[He et Young, 2006] Y. HE et S. YOUNG (2006). « Spoken language understanding using hidden state vector model ». *Speech Communication*, 48:262–275. 27

[Hermansky, 1990] Hynek HERMANSKY (1990). « Perceptual linear predictive (plp) analysis of speech ». *Acoustical Society of America Journal*, 87:1738–1752. 59

[Huang et Cox, 2003] Q. HUANG et S.J. COX (2003). « Automatic Call-Routing Without Transcriptions ». *Eighth European Conference on Speech Communication and Technology*. 78

[Jackendoff, 2002] R. JACKENDOFF (2002). *Foundations of languages*. UK : Oxford Univ. Press. 22

[Jamoussi *et al.*, 2005] S. JAMOUSSI, K. SMAILI et Jean-Paul HATON (2005). « From speech to sql queries : a speech understanding system ». *Conference on Artificial Intelligence workhop on spoken language understanding*. 18

[Katz, 1987] Slava M. KATZ (1987). « Estimation of probabilities from sparse data for the language model component of a speech recognizer ». *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401. 14

[Kuhn et De Mori, 1990] Roland KUHN et Renato DE MORI (1990). « A cache-based natural language method for speech recognition ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–582. 74

[Lavecchia *et al.*, 2007] Caroline LAVECCHIA, Kamel SMAÏLI, David LANGLOIS et Jean-Paul HATON (2007). « Using inter-lingual triggers for machine translation ». *INTERSPEECH*. 74

[Lavelle *et al.*, 1998] Alexia LAVELLE, Martine DE CALMÈS et Guy PÉRENNOU (1998). « Users's Behaviours in spontaneous Oral Dialogue Strategy Design ». *First Workshop on Text, Speech, Dialogue - TSD'98 , Brno, Czech Republic, 23/09/98-26/09/98*, pages 403–408, Masaryk University, Brno. Petr Sojka, Vaclav Matousek, Karel Pala, Ivan Kopecek. 20

[Levin et Pieraccini, 1995] Esther LEVIN et Roberto PIERACCINI (1995). « Concept-based spontaneous speech understanding system ». *EUROSPEECH*, pages 555–558, Madrid, Spain. 43

[Mariani, 1990] Joseph MARIANI (1990). « Reconnaissance de la parole : progrès et tendances ». *Traitement du signal*, 7(4):239–266. 10

[M.Benzequida *et al.*, 2006] M.BENZEQUIDA, Renato DE MORI, O. DEROO, S. DUPONT, T.ERBES, D. JOUVET, L. FISSORE, P. LAFACE, A. MERTINS, C. RIS, R. ROSE, V. TYAGI et C. WELLEKENS (2006). « Automatic speech recognition and intrinsic speech variation ». *IEEE ICASSP*. 21, 58

[McTear, 2006] M. MCTEAR (2006). « Spoken langage understanding for conversationnal dialog systems ». *IEEE/ACL Workshop on spoken language technology*, page 6, Palm Beach, Aruba. 17

[Miller *et al.*, 1994] Scott MILLER, Robert BOBROW, Richanrd SCHWARTZ et Robert INGRIA (1994). « Statistical language processing usign hidden understanding models ». *ACL*. 2, 26, 43

[Mohri *et al.*, 2002] Mehryar MOHRI, Fernando PEREIRA et Michael RILEY (2002). « Weighted finite-state transducers in speech recognition ». *Computer, Speech and Language*, 16(1):69–88. 48, 64, 71

[Moschitti *et al.*, 2007] Alessandro MOSCHITTI, Giuseppe RICCARDI et Christian RAYMOND (2007). « Spoken language understanding with kernels for syntactic/semantic structures ». *ASRU*, Kyoto, Japan. 28

[Ney *et al.*, 1992] Hermann NEY, Dieter MERGEL, Andreas NOLL et Annedore PAESELER (1992). « Data driven search organization for continuous speech recognition ». *IEEE Transactions on Signal Processing*, 40(2):272–281. 16

[Nocera *et al.*, 2002] P. NOCERA, G. LINARES et D. MASSONIE (2002). « Principes et performances du décodeur parole continue Speeral ». *Proc. Journées d'Etude sur la Parole (JEP)*. 42, 55

[Oger *et al.*, 2008] Stanislas OGER, Georges LINARÈS et Frédéric Béchet PASCAL NOCERA, dans (2008). « On-demand new word learning using worl wide web ». *ICASSP*. 96

[Papieni *et al.*, 1998] K.A. PAPIENI, S. ROUKOS et R.T. WARD (1998). « Maximum likelihood and discriminative training of direct translation models ». *IEEE ICASSP*, pages 189–192, Seattle, WA. 29

[Pieraccini *et al.*, 1991] Roberto PIERACCINI, Esther LEVIN et Chin-Hui LEE (1991). « Stochastic representation of conceptual structure in the atis task ». *Speech and Natural Language Workshop*, pages 121–124, Morgan Kaufmann publ, Los Altos, CA. 2, 26

[Ramshaw et Marcus, 1995] Lance RAMSHAW et Mitch MARCUS (1995). « Text chunking using transformation-based learning ». David YAROVSKY et Kenneth CHURCH, éditeurs : *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey. Association for Computational Linguistics. 46

[Raymond *et al.*, 2006] Christian RAYMOND, Frédéric BÉCHET, Renato DE MORI et Géraldine DAMNATI (2006). « On the use of finite state transducers for semantic interpretation ». *Speech Communication*, 48,3-4:288–304. 42

[Raymond *et al.*, 2007] Christian RAYMOND, Frédéric BÉCHET, Nathalie CAMELIN, Renato DE MORI et Géraldine DAMNATI (2007). « Sequential decision strategies for machine interpretation of speech ». *IEEE*, 15(1):162–171. 28

[Raymond et Riccardi, 2007] Christian RAYMOND et Giuseppe RICCARDI (2007). « Generative and discriminative algorithms for spoken language understanding ». *Interspeech*, Antwerp, Belgium. 29, 30

[Roark, 2002] Brian ROARK (2002). « Markov parsing : lattice rescoring with a statistical parser ». *Proceedings of the 40th ACL meeting, Philadelphia*. 44

[Rosset *et al.*, 2006] Sophie ROSSET, Olivier GALIBERT, Gabriel ILLOUZ et Aurélien MAX (2006). « Intégrating spoken dialog and question answering : The ritel project ». *INTERSPEECH'06*. 87

[Rosset et Tribout, 2005] S. ROSSET et D. TRIBOUT (2005). « Multi-level information and automatic dialog acts detection in human-human spoken dialogs ». *EUROSPEECH'05*, Lisbon. 37

[Sarikaya *et al.*, 2005] R. SARIKAYA, Y. GAO, M. PICHENY et H. ERDOGAN (2005). « Semantic confidence measurement for spoken dialog systems ». *In IEEE Transactions on Speech and Audio Processing*, 13(4):534–545. 31

[Savoy, 2006] Jacues SAVOY (2006). « Un regard statistique sur l'évaluation de performance : l'exemple clef 2005 ». *CORIA*. 89

[Schapire et Singer, 2000] Robert E. SCHAPIRE et Yoram SINGER (2000). « BoosTexter : A boosting-based system for text categorization ». *Machine Learning*, 39:135–168. 90

[Seneff, 1989] Stéphanie SENEFF (1989). « Tina : A probabilistic syntactic parser for speech understanding systems ». *Proc. of the Speech and Natural Language Workshop*, pages 168–178, Philadelphia, PA, USA. 26

[Servan *et al.*, 2006] Christophe SERVAN, Christian RAYMOND, Frédéric BÉCHET et Pascal NOCÉRA (2006). « Conceptual decoding from word lattices : application to the spoken coprus media ». *INTERSPEECH - ICSLP*, page 4, Pittsburgh (USA). 59

[Shapire *et al.*, 2005] R.E. SHAPIRE, M. ROCHERY, M. RAHIN et N. GUPTA (2005). « Boosting with prior knowledge for call classification ». *IEEE Transaction Speech Audio Processing*, 13(2):174–182. 28

[Smaïli *et al.*, 1999] Kamel SMAÏLI, Imed ZITOUNI et Jean-Paul HATON (1999). « Towards a better collaboration between a n-class and a n-gram language model ». *International Workshop on speech and communication*, Moscou. 12

[Spérandio et Letang-Figeac, 1986] J.C. SPÉRANDIO et C. LETANG-FIGEAC (1986). « Simulation expérimentale de dialogues oraux en communication homme-machine ». Communication parlée, GRECO-CNRS, Nancy, France. 20

[Tur *et al.*, 2005] G. TUR, D. Hakkani TÜR et A. CHOMTIMONGKOL (2005). « Semi-supervised learning for sopken language understanding using semantic role labeling ». *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 232–237, Puerto Rico. 23

[Villaneau *et al.*, 2004] J. VILLANEAU, J.-Y. ANTOINE et O. RIDOUX (2004). « Logical approach to natural language understanding in a spoken dialogue system ». *Text, Speech and Dialogue, 7th International Conference*. 37

[Wang *et al.*, 2005] Y. WANG, L. DENG et A. ACERO (Sep. 2005). « Spoken language understanding ». *IEEE Signal Processing Magazine*, 22(5):16–31. 26, 27

[Wang *et al.*, 2002] Ye-Yi WANG, A. ACERO, C. CHELBA, B. FREY et L. WONG (2002). « Combination of statistical and rule-based approaches for spoken language understanding ». *Proc. International Conference on Spoken Language Processing*, pages 609–613, Denver, CO, USA. 43, 44

[W.A.Wood, 1975] W.A.WOOD (1975). *Representation and Understanding*, chapitre What is a link ? D.G. Bobrw and A. Collins. 17

[Zitouni *et al.*, 2003] Imed ZITOUNI, H.-K. JEFF KUO et C.-H. LEE (2003). « Boosting and combination of classifiers for natural language call routing systems ». *Speech Communication*, 41(4):647–661. 78

[Zitouni *et al.*, 1999] Imed ZITOUNI, Jean-François MARI, Kamel SMAÏLI et Jean-Paul HATON (1999). « Variable-length sequence language model for large vocabulary continuous dictation machine : The n-seqgram approach ». *European Conference on Speech Communication and Technology*, volume 4, pages 1811–1814, Budapest. 16

Publications Personnelles

[Barrault *et al.*, 2008] Loïc BARRAULT, Christophe SERVAN, Driss MATROUF, Georges LINARÈS et Renato De MORI (2008). « Frame-based acoustic feature integration for speech understanding ». *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'08)*, Las Vegas. 58

[Bonneau-Maynard *et al.*, 2006] Hélene BONNEAU-MAYNARD, Christelle AYACHE, Frédéric BÉCHET, Alexandre DENIS, Anne KHUN, Fabrice LEFEVRE, Djamel MOSTEFA, Mathieu QUINARD, Sophie ROSSET, Christophe SERVAN et Jeanne VILLANEAU (2006). « Results of french evalda-media evaluation campaign for litteral understanding ». *Language Resources and Evaluation Conference (LREC'06)*, pages 2054–2059, Gènes (Italie). 38, 62

[Duvert *et al.*, 2008a] Frédéric DUVERT, Marie-Jean MEURS, Christophe SERVAN, Frédéric BÉCHET, Fabrice LEFÈVRE et Renato De MORI (2008a). « Composition sémantique pour la compréhension de la parole dans le cadre de dialogue ». *Journées d'Études sur la Parole (JEP)*, Avignon.

[Duvert *et al.*, 2008b] Frédéric DUVERT, Marie-Jean MEURS, Christophe SERVAN, Frédéric BÉCHET, Fabrice LEFÈVRE et Renato De MORI (2008b). « Semantic composition process in a speech understanding system ». *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'08)*, Las Vegas.

[Servan, 2006] Christophe SERVAN (2006). « Utilisation des transducteurs dans le décodage conceptuel : application au corpus media ». *MAJECSTIC*, Lorient.

[Servan et Bechet, 2006] Christophe SERVAN et Frederic BECHET (2006). « Décodage conceptuel et apprentissage automatique : application au corpus de dialogue homme-machine media ». *TALN*, Leuven (Belgique).

[Servan et Béchet, 2008] Christophe SERVAN et Frédéric BÉCHET (2008). « Fast-call classification system development without in-domain training data ». *INTERSPEECH*, Brisbane (Australie).

[Servan *et al.*, 2006a] Christophe SERVAN, Christian RAYMOND, Frédéric BÉCHET et Pascal NOCÉRA (2006a). « Conceptual decoding from word lattices : application to the spoken coprus media ». *INTERSPEECH - ICSLP*, page 4, Pittsburgh (USA). 59

[Servan *et al.*, 2006b] Christophe SERVAN, Christian RAYMOND, Frédéric BÉCHET et Pascal NOCÉRA (2006b). « Décodage conceptuel à partir de graphes de mots sur le corpus de dialogue homme-machine media ». *Journées d'Études sur la Parole (JEP)*, Dinard.