Cadre expérimental

Sommaire

Ce chapitre introduit le cadre expérimental défini par l'utilisation du service 3000 ainsi que les différentes problématiques liées à l'utilisation des corpus réels. Le chapitre est divisé en cinq sections. Dans la première section 4.1, nous faisons une analyse des énoncés utilisateur et nous posons les problématiques liées au comportement du système sur différents types d'énoncés. Dans la section 4.2, nous présentons les modèles de langage et acoustique utilisés par le SRAP ainsi que le lexique de l'application. Une description des données expérimentales (corpus d'apprentissage, de développement et de test) est réalisée dans la section 4.3. La section 4.4 présente les problèmes liées à l'évaluation au niveau mot des corpus réels et décrit plusieurs méthodes de normalisation des données. La dernière section 4.5 est consacrée à l'évaluation au niveau interprétation sur le corpus de test.

4.1 Caractérisation des énoncés utilisateur

Pour le service 3000 on distingue deux types de dialogues. Certains utilisateurs appellent afin d'activer des services auxquels ils ont déjà souscrit, comme le nombre de minutes utilisées ou le transfert de leur numéro de téléphone. Pour ce type de demande l'utilisateur est, ensuite, transféré vers un service vocal dédié à ces tâches spécifiques. Dans ce cas, le service 3000 peut être vu comme un service de routage qui redirige,

de manière efficace, les utilisateurs, en exploitant éventuellement leur profil utilisateur pour des informations complémentaires. Ces utilisateurs sont assez familiers avec le système et l'utilisent régulièrement. Ils sont plus prédisposés à utiliser des phrases courtes ou seulement des mots clés et l'interaction avec le système est rapide (entre deux et trois tours de parole avant d'être redirigés vers le service demandé). Ce type de dialogue représente 80% des appels sur le service 3000. Pour les 20% restants, l'interaction avec l'utilisateur est entièrement gérée par le service 3000 lui même. Ces dialogues proviennent généralement des utilisateurs demandant des informations sur certains services ou des utilisateurs qui cherchent à s'abonner à un nouveau service. Pour ce type de dialogue, la longueur moyenne de la phrase est plus élevée ainsi que le nombre moyen de tours de parole. Les utilisateurs sont moins familiers avec l'application et le nombre moyen de disfluences ainsi que de mots hors vocabulaire (les mots qui ne sont pas couverts par le lexique de l'application) augmente.
Un autre aspect important pour ce deuxième type de dialogue est le taux assez élevé de commentaires de la part des utilisateurs. Un commentaire est une phrase prononcée par l'utilisateur qui est considérée comme étant hors du domaine de l'application. On distingue plusieurs catégories de commentaires. L'utilisateur peut se parler à lui même ("*oh non, j'en ai marre de ce truc*"), peut être surpris ("*qu'est-ce que je dois dire maintenant ?*"), peut être énervé (*j'ai déjà dit ça*") ou peut même insulter le système. On distingue aussi un type de commentaire que nous appelons "*aparté*", lorsque l'utilisateur parle à une tierce personne se trouvant à coté de lui ("*va ranger ta chambre*"). Les commentaires peuvent constituer une phrase entière mais il existe aussi des commentaires qui peuvent se retrouver avec de l'information utile dans le même énoncé.
Le problème se pose lorsque le système essaie de décoder et d'interpréter ces phrases. Comme elles sont hors du domaine de l'application, une partie des mots utilisés sont des mots hors vocabulaire, mais plus important, le module de compréhension n'est pas capable de les interpréter. Dans le meilleur des cas, une phrase qui est un commentaire génère une incompréhension de la part du système qui demandera à l'utilisateur de répéter. Mais il peut y avoir aussi le cas opposé, où une telle phrase génère une interprétation (les mots reconnus peuvent allumer des concepts auxquels correspond une règle d'interprétation) ce qui peut conduire le DM dans une mauvaise direction. Ce dernier cas oblige l'utilisateur à revenir en arrière et à répéter sa demande et donne l'impression que le système a commis une erreur, ce qui n'est pas le cas. Il existe aussi la possibilité qu'une phrase contenant des commentaires mais aussi de l'information utile ne puisse pas être interprétée par le système qui génère une réponse d'incompréhension. Dans tous ces cas, un autre problème majeur outre le comportement du système est le traitement de la phrase prononcée par l'utilisateur. Nous allons montrer que ces traitements peuvent être très complexes et coûteux en termes de temps de calcul et de ressources utilisées et nous allons proposer, dans le chapitre 7, différentes stratégies afin de contourner ce problème.
Afin de mieux analyser un système de compréhension de la parole, il est important de ne pas se contenter d'évaluer les performances du système dans son ensemble mais plutôt d'observer son comportement sur les différents types de messages auxquels il est confronté. La première grande distinction concerne les messages qui contiennent une interprétation valide et ceux qui n'en contiennent pas. Les messages qui ne contiennent

pas d'interprétation valide sont considérés, du point de vue du système, comme étant à rejeter. Une deuxième grande distinction concerne la présence des commentaires. Puisque les commentaires ne sont pas couverts par le domaine de l'application nous allons appeler les séquences de mots qui forment les commentaires, de la parole *Hors-Domaine*, et par conséquence tous ce qui n'est pas un commentaire est appelé parole *Dans-le-Domaine*. La parole *Hors-Domaine*, de par sa définition, ne peut pas donner lieu à une interprétation valide et fait donc partie des énoncés à rejeter. Suite à ces observations nous pouvons distinguer trois catégories de messages, dont les deux premières sont des messages à rejeter :

- **C1-Non-Parole**- Les énoncés non-parole. Tous les énoncés ne contenant que du bruit, admis à tort par le module de détection Bruit/Parole. Ce sont des énoncés qui ne donnent lieu à aucune interprétation valide et sont donc à rejeter.
- **C2-Hors-Domaine**- La parole *Hors-Domaine*. Tous les énoncés qui contiennent des commentaires.
- **C3-Dans-le-Domaine**- La parole *Dans-le-Domaine*. Tous les énoncés qui ne rentrent pas dans le deux premières catégories.

4.2 Modèles utilisés par le service 3000

Le système de reconnaissance de la parole (SRAP) du service 3000 utilise un lexique de 2548 mots. Comme nous l'avons expliqué au 3.3, le lexique est formé de deux types de mots : les mots **non-vides**, qui sont en proportion de 44% dans le lexique et les mots **vides** qui constituent le reste du lexique.
Comme expliqué au 3.1, le SRAP est précédé par un module de détection Bruit/Parole. Ce module à pour but d'éviter les activations intempestives du SRAP dues aux bruits environnants. Afin de détecter et rejeter les énoncés ne contenant que du bruit qui n'ont pas été détectés par le module Bruit/Parole, le SRAP utilise un sous-modèle de détection (rejet) des bruits. Afin de détecter les mots hors-vocabulaire (*OOV - Out of Vocabulary*), un modèle phonétique contraint par la longueur (Hamimed et Damnati, 2002) est utilisé. Le SRAP utilise également un modèle phonétique avec des phonèmes hors-contexte afin de détecter les faux départs, que nous notons *SPR* (*Speech Repair*) dans ce document.
Le modèle de langage utilisé est un modèle de type bi-classes dont la majorité des classes sont constituées d'un seul mot (voir 1.3.2 pour la description de ce type de modèle). Afin de détecter les commentaires, un sous-modèle de langage (Damnati et al., 2007) est inclut dans le modèle principal. Pour construire ce sous-modèle, les commentaires ont été annotés manuellement sur un corpus d'apprentissage afin de pouvoir séparer les segments *Hors-Domaine* des segments *Dans-le-Domaine*. Les commentaires de type *aparté* (voir la section 4.1) peuvent être très diversifiés et ne sont pas modélisés par ce modèle de langage. Les autres commentaires qui sont adressés au système ont tendance à être assez redondants et donc plus facile à modéliser. Le corpus d'apprentissage pour le sous-modèle de détection de la parole *Hors-Domaine* contient 1712 séquences ayant de la parole *Hors-Domaine* et le lexique utilisé est de 765 mots.
Le sous-modèle, appelé LM^{HD} a été intégré au modèle général LM^{G}. Dans le corpus

d'apprentissage du modèle général (voir la section 4.3), les commentaires ont été annotés et remplacés par l'étiquette <*COMMENTAIRE*>. Cette étiquette est rajoutée ensuite dans le bigramme général et les probabilités $P(< COMMENTAIRE > |w)$ et $P(w| < COMMENTAIRE >)$ sont apprises en même temps que les autres probabilités du bigramme (suivant le principe des classes *a priori*). Pendant le décodage, les probabilités du modèle général et celles du sous-modèle LM^{HD} sont combinées (voir 1.3.3 pour plus de détails sur l'imbrication des modèles). Pour certaines évaluations, comme celles présentées dans les chapitres 5 et 6, le sous-modèle de détection de commentaires n'est pas intégré au modèle général. Les mots ou séquences de mots qui auraient été détectées par ce sous-modèle sont détectés alors comme tout autre mot par le SRAP en utilisant seulement le modèle de langage général. Ces mots ou séquences de mots ne sont donc pas idéntifiés comme un commentaire.

4.3 Description des données expérimentales

L'avantage d'utiliser un service déployé comme le 3000 est de pouvoir travailler sur des énoncés issus des utilisateurs qui se trouvent en situation réelle d'utilisation du système. Tous les corpus utilisés sont formés des énoncés issus des utilisateurs faisant appel au service et collectés sur différentes périodes de temps.

Données d'apprentissage

Les données d'apprentissage du modèle de langage se présentent sous la forme d'un corpus constitué de transcriptions de phrases prononcées par les utilisateurs de l'application. Le corpus d'apprentissage est constitué de 44000 énoncés.
Ce corpus a été utilisé pour l'apprentissage du bigramme mais aussi pour l'apprentissage du sous-modèle de détection des commentaires. Pour ce dernier, les corpus a été annoté manuellement pour différentier les commentaires du reste de l'énoncé.
Par la suite, le corpus d'apprentissage sera noté **App**.

Données de développement et de test

Plusieurs corpus constituent les données de développement et de test. Chaque énoncé constituant les corpus est associé à une phrase de référence qui correspond à la transcription manuelle des paroles prononcées par l'utilisateur. Dans cette référence les commentaires peuvent être ou ne pas être annotés. Outre la meilleure hypothèse de reconnaissance générée par le SRAP pour chaque énoncé, un graphe de mots est également généré. Les scores acoustiques associés aux mots du graphe sont calculés lors de cette première passe de décodage.
Plusieurs corpus sont utilisés dans le cadre de nos travaux :

- **Dev** est un corpus de développement, constitué de 1764 énoncé, ce qui représente 905 dialogues. Les phrases de référence associées à ces énoncés sont constituées de 4843 mots.
- **Test_I** est un premier corpus de test constitué de 2005 enregistrements. Les transcriptions de référence contiennent un total de 5595 mots.
- **Test_II** est un second corpus de test constitué de 6501 énoncés. Les transcriptions de référence contiennent un total de 13890 mots.

Le corpus **Dev** est utilisé principalement pour le calibrage des coefficients de la régression logistique (expliqué au 2.2.4). Le corpus **Test_I** est utilisé pour effectuer la comparaison des performances en termes de *WER* de deux algorithmes de génération de réseaux de confusion présentés au chapitre 5. Le corpus **Test_II** est utilisé dans la majorité des évaluations effectuées dans ces travaux.

Corpus	# Énonces	# Total de mots dans la référence	Longueur moyenne
Dev	1764	4843	2.7
Test_I	2005	5595	2.8
Test_II	6501	13890	2.1

TABLE *4.1* – *Description des données de développement et de test*

Le tableau 4.1 donne le détail de chaque corpus en termes de nombre total d'énoncés, nombre total de mots dans la référence et la longueur moyenne de chaque référence (le nombre moyen de mots par énoncé de référence). Le tableau 4.2 donne la taille des graphes de mots pour les deux corpus de test en termes de nombre de transitions par graphe et de nombre de transitions par mot de la référence (calculé comme une moyenne sur le nombre de transitions par mot de la référence calculé pour chaque enregistrement). Plusieurs transitions dans le graphe peuvent porter la même hypothèse de mot. On observe que la taille des graphes est assez grande par rapport au nombre moyen de mots dans la référence. Par ailleurs, pour **Test_II**, le nombre moyen d'hypothèses de mots différent présents dans le graphe est de 42 mots.

Corpus	# Graphes de mots	# Transitions / Graphe	# Transitions / Mot de la référence
Test_I	2005	26800	13600
Test_II	6501	17000	12000

TABLE 4.2 – *Statistiques des graphes de mots construits sur les corpus de test*

Les graphes de mots sur l'ensemble des corpus sont générés en utilisant le modèle de langage décrit au 1.3.3. Le sous-modèle de détection de commentaires n'est toutefois pas inclus dans le modèle général. Ceci est du au fait qu'il est difficile d'estimer la probabilité *a posteriori* sur les segments *Hors-Domaine*. En effet, si on utilise le sous-modèle de détection des commentaires, dans le graphe de mots les séquences de mots détectées lors du décodage comme étant des commentaires seraient remplacées par l'étiquette <*COMMENTAIRE*>. Toutefois, dans le graphe de mots, l'étiquette <*COMMENTAIRE*> n'a pas une vraisemblance acoustique associée et donc le calcul de sa probabilité *a posteriori* passe d'abord par un calcul des probabilités *a posteriori* de chaque mot ayant

été détecté comme faisant partie du commentaire. L'algorithme doit alors avoir accès à la séquence de mots détectée et aussi aux vraisemblances acoustiques de chaque mot. Même en ayant calculé ces probabilités, l'obtention de la probabilité *a posteriori* du commentaire n'est pas aisée ; une simple multiplication des probabilités *a posteriori* des mots n'est pas possible parce que on compterait plusieurs fois les mêmes chemins dans le calcul de la probabilité *a posteriori* du commentaire (la probabilité *a posteriori* d'une transition est la somme des probabilités *a posteriori* des chemins passant par cette transition normalisé par la somme de tous les chemins dans le graphe). La complexité de calcul étant trop importante nous avons décidé de générer les graphes de mots sans détection des commentaires.

4.4 Evaluation du WER sur un corpus réel

Du point du vu du système, potentiellement tout mot en entrée est un mot à reconnaître. Le calcul du *WER* se fait en évaluant la distance de Levensthein (voir 1.6) entre la référence et l'hypothèse de reconnaissance fournie par le système. Toutefois, comme nous l'avons décrit, lorsqu'on travaille avec un corpus constitué d'énoncés de parole continue collectés à partir d'une application déployée on doit tenir compte de multiples événements.
Prenons l'exemple d'un énoncé non-parole dans le tableau 4.3. Un tel énoncé est à rejeter, donc lors de la transcription manuelle l'étiquette <*REJET*> lui sera attribuée et la référence ne contiendra que cette étiquette. Lors du décodage, un certain nombre de ces énoncés sont reconnus par le système comme des énoncés non-parole (ne contenant que du bruit) grâce au sous-modèle de rejet des bruits et, par conséquence, vont se voir attribués l'étiquette <*REJET*>. Par contre il est aussi possible que le système fournisse une hypothèse vide (le décodage n'est pas passé par le sous-modèle de rejet des bruits et aucun mot n'a été reconnu) ou une hypothèse contenant l'étiquette *OOV*(le décodage est passé par le modèle de détection des mots hors-vocabulaire).
Dans le premier cas, lors du calcul du *WER*, aucune erreur ne sera comptée mais pour le deuxième cas, l'algorithme de calcul va compter une omission (l'étiquette <*REJET*> dans la référence a été omise dans l'hypothèse de reconnaissance) ou une substitution (l'étiquette <*REJET*> dans la référence a été substituée avec l'étiquette *OOV* dans l'hypothèse de reconnaissance) dans le dernier cas.

Type d'énoncé	Référence	Hypothèse de reconnaissance possibles	# Erreurs comptés dans calcul WER
Énonce non-parole	"<*REJET*>"	"<*REJET*>"	0
		""	1 omission
		"*OOV*"	1 substitution

TABLE 4.3 – *Exemple d'annotation dans le calcul du WER*

Il existe des situations plus complexes qui peuvent donner un nombre plus important d'erreurs. Par exemple, un *aparté* (un type d'énoncé *Hors-Domaine* difficile à modéliser de part sa diversité) est un énoncé à rejeter et la référence est constituée seulement de

l'étiquette <*REJET*>. Par contre, lors du décodage, le système peut détecter, entre autre, un enchaînement de mots hors-vocabulaire et l'hypothèse fournie sera constituée de plusieurs étiquettes *OOV*. Donc, lors du calcul du *WER*, une substitution et plusieurs insertions sont comptées.
La façon d'annoter tous ces événements influe directement sur le calcul du *WER*. L'annotation doit être tout d'abord homogène ; par exemple, l'étiquette attribuée à un énoncé non-parole dans la référence doit être la même que celle attribué par le système suite à une détection non-parole faite par le sous-modèle de rejet des bruits. Le problème qui se pose ensuite et de savoir si certains événements ayant des annotations différentes n'ont pas la même signification dans le contexte plus large du système de dialogue. Par exemple, l'hypothèse de reconnaissance vide qui génère une omission dans le tableau 4.3 peut être considérée comme un énoncé rejeté par le système de dialogue. Ceci implique que l'hypothèse de reconnaissance contient désormais l'étiquette <*REJET*> et aucune erreur n'est comptée dans le calcul du *WER*. On peut aussi donner la même signification à l'hypothèse de reconnaissance contenant l'étiquette *OOV*. On considère alors que le système rejette un énoncé pour lequel il ne détecte que des mots hors vocabulaire (voir des *SPR*).
Il est évident que pour réaliser une évaluation pertinente du *WER* sur un corpus réel, dans le contexte plus large du système de dialogue, nous avons besoin d'une annotation homogène et de réaliser une normalisation de la notion du rejet du aux différents événements comme les détections non-parole, les *OOV*, les *SPR*, les commentaires.

4.4.1 Méthodes de normalisation

Nous définissons quatre méthodes de normalisation des énoncés et nous montrons leur influence sur le calcul du *WER* à travers un exemple. Pour définir la première méthode, nous partons de la prémisse que tout mot en entrée est un mot à reconnaître et qu'en l'absence d'un mot aucune étiquette ne sera attribuée à un énoncé (par exemple, les énoncés non-parole ne se verront pas attribuer l'étiquette <*REJET*>). Dans la deuxième méthode, les énoncés non-parole et les énoncés ayant une référence vide sont considérés comme étant des énoncés à rejeter et leur référence est remplacée par l'étiquette <*REJET*>. Le SRAP attribue la même étiquette si l'hypothèse de reconnaissance est vide ou si un énoncé est détecté par le sous-modèle de rejet des bruits. La troisième méthode introduit la notion de rejet énoncé parole (les énoncés qui ne contiennent que des *OOV* ou des *SPR*). Les énoncés considérés comme étant des rejets énoncé parole se voient attribuer l'étiquette <*REJET*>. Dans ces trois premières méthodes la détection de commentaires n'est pas faite (le sous-modèle de langage n'est pas inclut dans le modèle général lors du décodage). De ce fait, l'annotation des commentaires dans la référence est ignorée et les mots ayant été annotés en commentaire sont traités comme tout autre mot de l'application. La dernière méthode est utilisée dans le cas où le système peut détecter les commentaires et, par conséquence, les séquences de mots annotées en commentaire dans la référence sont remplacées par l'étiquette <*COMMENTAIRE*>. La notion de rejet énoncé parole est alors étendue aux commentaires.
Afin de mieux illustrer l'effet de ces méthodes et l'importance de la normalisation de la référence et de l'hypothèse de reconnaissance dans le calcul du taux d'erreur mot

(*WER*), nous prenons l'exemple du tableau 4.4. Il est constitué de quatre énoncés réels qui résument les principaux problèmes rencontrés dans le calcul du WER sur un corpus réel.

Énonce	Transcription manuelle annotée	Hypothèse de reconnaissance fournie
(f)acture	*vide*	*rejet* [1]
énonce non-parole	*bruit*	*OOV*
faux départ	*SPR*	*OOV OOV*
oh merde oh là oh	[com :] oh merde oh là oh [com :]	[com :] ah merde oh là c' est trop [:com] [2]
oh là là payer ma facture	[com :] oh là là [:com] payer ma facture	oh payer ma facture

TABLE 4.4 – *Exemple d'énoncés réels*

Dans cet exemple, le premier énoncé contient un mot tronqué (le "f" est prononcé, mais le signal de parole est coupé par l'automate de détection bruit/parole). Pour le SRAP il est difficile de reconnaître ce mot à partir du signal tronqué et donc la transcription manuelle est vide (même si en écoutant l'énoncé tronqué, un humain est capable de reconnaître le mot prononcé). Les deux exemples suivants correspondent respectivement à un énoncé non-parole (du bruit) et à un faux départ (*SPR*). L'exemple suivant contient de la parole qui est en fait un commentaire, d'où les deux étiquettes ([com :], [:com]) dans la transcription annotée. On retrouve les mêmes étiquettes dans l'hypothèse de reconnaissance parce que le décodage est passé par le sous-modèle de détection des commentaires. La même séquence de mots est toutefois détectée par le SRAP sans utiliser le sous-modèle de détection des commentaires (dans ce cas les étiquettes ne sont pas rajoutées). Le dernier exemple contient de la parole *Hors-Domaine* et de la parole *Dans-le-Domaine*, d'où l'annotation dans la transcription manuelle d'une partie de l'énoncé. L'hypothèse de reconnaissance ne contient aucune annotation car aucun commentaire n'est détecté par le SRAP utilisant le sous-modèle de langage.
Les tableaux qui suivent pour chaque méthode montrent les différences dans le calcul du *WER* en fonction de la normalisation effectuée sur les énoncées de l'exemple du tableau 4.4. Chaque tableau donne la référence obtenue par la normalisation de l'annotation manuelle ainsi que de l'hypothèse de reconnaissance. La dernière partie de chaque tableau présente le détail du nombre d'erreurs et de mots corrects sur quatre colonnes. La première colonne "*C*", donne le nombre de mots corrects. Les trois colonnes suivantes, dans l'ordre "*I*" "*S*" "*O*", donnent respectivement le nombre d'insertions, de substitutions et d'omissions. La dernière ligne de chaque tableau donne la valeur du *WER*.

Méthode 1 : Les énoncés non-parole ne sont pas étiquetés et la référence correspondant à un tel énoncé est vide. Il en est de même pour l'hypothèse de reconnaissance qui sera vide lorsque le système passe par le sous-modèle de rejet des bruits.

1. Le SRAP est passé par le modèle de rejet des bruits.
2. La même séquence de mots est également détectée par le SRAP sans utiliser le sous-modèle des commentaires.

Référence de l'énoncé	Hypothèse de reconnaissance	Détails erreurs			
		C	I	S	O
		0	0	0	0
	OOV	0	1	0	0
SPR	*OOV OOV*	0	1	1	0
oh merde oh là oh	ah merde oh là c' est trop	3	2	2	0
oh là là payer ma facture	oh payer ma facture	4	0	0	2
WER=75%		7	4	3	2

TABLE *4.5* – *Calcul du WER :* ***Méthode1***

Dans ce premier tableau les énoncés non-parole ne sont pas étiquetés et la référence est vide. Il en est de même pour l'hypothèse de reconnaissance du premier énoncé de l'exemple. Malgré le fait que le SRAP est passé par le sous-modèle de rejet des bruits, l'hypothèse est vide car aucune étiquette n'est attribué.

Méthode 2 : Les énoncés non-parole sont étiquetés et la référence correspondant à un tel énoncé contient l'étiquette <*REJET*>. Il en est de même pour l'hypothèse de reconnaissance qui contiendra l'étiquette <*REJET*> lorsque le système passe par le sous-modèle de rejet des bruits. Une référence d'un énoncé ou une hypothèse de reconnaissance vide sont également considérées comme un rejet et l'étiquette <*REJET*> leur est attribuée.

Référence de l'énoncé	Hypothèse de reconnaissance	Détails erreurs			
		C	I	S	O
<*REJET*>	<*REJET*>	1	0	0	0
<*REJET*>	*OOV*	0	0	1	0
SPR	*OOV OOV*	0	1	1	0
oh merde oh là oh	ah merde oh là c' est trop	3	2	2	0
oh là là payer ma facture	oh payer ma facture	4	0	0	2
WER=64%		8	3	4	2

TABLE *4.6* – *Calcul du WER :* ***Méthode2***

Les références et les hypothèses de reconnaissance qui étaient vides dans la première méthode contiennent l'étiquette <*REJET*>. On observe que le nombre total d'erreurs est le même, car l'insertion dans le deuxième énoncé est devenue une substitution du fait de la modification de la référence. La valeur plus petite du *WER* s'explique par un nombre plus grand de mots dans la référence (on passe de 12 à 14 mots) qui a pour résultat un dénominateur plus grand pour un numérateur constant dans la formule de calcul du *WER*.

Méthode 3 : Rejet énoncé parole. En plus des conditions de la **Méthode2**, un énoncé qui ne contient que des mots hors-vocabulaire *OOV* ou des *SPR* est compté comme un énoncé à rejeter (rejet énoncé parole). La référence et l'hypothèse de reconnaissance correspondantes sont alors remplacées par l'étiquette <*REJET*>.

L'utilisation de la **Méthode3** introduit la notion de rejet énoncé parole. Le deuxième énoncé est considéré par le système comme étant un rejet et annoté en conséquence. Le troisième énoncé devient un énoncé à rejeter et il est aussi considéré par le système

Référence de l'énoncé	Hypothèse de reconnaissance	Détails erreurs			
		C	I	S	O
<*REJET*>	<*REJET*>	1	0	0	0
<*REJET*>	<*REJET*>	1	0	0	0
<*REJET*>	<*REJET*>	1	0	0	0
oh merde oh là oh	ah merde oh là c' est trop	3	2	2	0
oh là là payer ma facture	oh payer ma facture	4	0	0	2
WER=43%		10	2	2	2

TABLE *4.7* – *Calcul du WER :* ***Méthode3***

comme un rejet. Cette fois ci, la diminution de la valeur du WER n'est pas due à un nombre plus grand de mots dans la référence (il est le même que pour la précédente) mais bien à un nombre plus petit d'erreurs. Cette baisse du nombre d'erreurs est accompagnée par une augmentation du nombre de mots correctement reconnus.

Méthode 4 : Rejet énoncé parole et détection des commentaires. Les commentaires annotés dans la référence sont remplacées par l'étiquette <*COMMENTAIRE*>. Il en est de même pour les commentaires détectés par le SRAP lors du décodage. La notion de rejet énoncé parole définie à la **Méthode3**, est étendue aux commentaires (un énoncé qui ne contient que des *OOV*, des *SPR* ou des commentaires est compté comme un rejet et remplacé par l'étiquette <*REJET*>).

Référence de l'énoncé	Hypothèse de reconnaissance	Détails erreurs			
		C	I	S	O
<*REJET*>	<*REJET*>	1	0	0	0
<*REJET*>	<*REJET*>	1	0	0	0
<*REJET*>	<*REJET*>	1	0	0	0
<*REJET*>	<*REJET*>	1	0	0	0
<*COMMENTAIRE*>payer ma facture	oh payer ma facture	3	0	1	0
WER=12%		7	0	1	0

TABLE *4.8* – *Calcul du WER :* ***Méthode4***

Le fait d'avoir annoté le quatrième énoncé comme étant un commentaire et sa détection en tant que commentaire font que les séquences de mots sont remplacés par l'étiquette <*REJET*> dans la référence et dans l'hypothèse. En revanche, dans le dernier énoncé, seul la séquence de mots annotée est remplacée par l'étiquette <*COMMENTAIRE*>. L'énoncé n'est pas considéré comme un rejet énoncé parole car en plus du commentaire il contient d'autres mots. Pour cet énoncé le SRAP n'a détecté aucun commentaire.

Comme montré à travers ces exemples, le calcul du *WER* sur un corpus réel n'est pas aisé et une normalisation de la référence ainsi que de l'hypothèse de reconnaissance est nécessaire afin d'effectuer une évaluation pertinente et homogène dans le contexte du système de dialogue. La différence entre les valeurs du *WER* obtenues avec les deux premières méthodes alors que le nombre d'erreurs est le même montre l'importance d'utiliser la même méthode de calcul pour pouvoir comparer correctement plusieurs évaluations (le nombre de mots de la référence diffère pour les deux méthodes).
Dans les méthodes que nous avons présentées, seules les deux dernières sont utilisées

dans les évaluations présentées dans ce rapport. Ainsi, la **Méthode3** est utilisée dans les évaluations où la détection des commentaires n'est pas réalisée et la **Méthode4** dans le cas contraire.

4.4.2 Evaluation du taux d'erreur mot

Dans cette section nous présentons les évaluations effectuées en termes de *WER* sur la *1-best* des deux corpus de test utilisant les quatre méthodes de normalisation présentées. Le tableau 4.9 présente l'évaluation en termes de *WER* sur le corpus de test **Test_I**. L'évaluation du *WER* pour le corpus de test **Test_II** est présenté dans le tableau 4.10. Dans les deux tableaux indique aussi le taux de mots correctement reconnus ("C") ainsi que les taux d'insertion, de substitution et d'omission ("I", "S", "O").

Test_I	WER	C	I	S	O
Methode 1	45.4%	67.0%	12.5%	24.5%	8.4%
Methode 2	45.1%	66.4%	11.4%	25.4%	8.3%
Methode 3	45.1%	66.5%	11.6%	25.2%	8.3%
Methode 4	44.2%	72.1%	16.3%	20.3%	7.6%

TABLE *4.9* – *Evaluation de la* 1-best *sur le corpus* ***Test_I*** *à l'aide des quatre méthodes de normalisation*

Test_II	WER	C	I	S	O
Methode 1	47.7%	74.4%	22.2%	18.8%	6.7%
Methode 2	43.0%	73.1%	16.1%	21.4%	5.5%
Methode 3	42.0%	73.9%	15.9%	20.6%	5.5%
Methode 4	41.8%	76.0%	17.8%	19.4%	4.6%

TABLE *4.10* – *Evaluation de la* 1-best *sur le corpus* ***Test_II*** *à l'aide des quatre méthodes de normalisation*

Nous rappelons que les valeurs des *WER* sur les quatre méthodes ne peuvent pas être comparées directement entre elles car la méthode utilisée pour la normalisation de la référence n'est pas la même et donc le nombre de mots dans la référence diffère d'une méthode à l'autre. La dernière ligne de chaque tableau correspond à l'utilisation du sous-modèle de détection des commentaires dans la première passe de reconnaissance. Pour les évaluations présentées sur les trois premières lignes, le sous-modèle de détection des commentaires n'est pas utilisé et donc les commentaires ne sont pas détectés. En conséquence, les annotations des commentaires dans la référence sont ignorées et la référence est constituée de la séquence de mots qui été annotée comme étant un commentaire.

4.5 Evaluation du taux d'erreur d'interprétation

Dans cette section nous présentons une évaluation de la *1-best* des corpus de test au niveau interprétation en calculant l'*IER* (voir 3.4 pour la définition de cette métrique).

Contrairement à l'évaluation au niveau mot, au niveau interprétation on n'a pas besoin d'effectuer une normalisation des données. Ceci est du au fait que les problèmes de normalisation présentée au 4.4 (exceptée la Méthode 4 qui introduit la détection de commantaires) sont résolus directement par le processus d'interprétation. Par exemple un énoncé vide sera interprété comme étant un rejet, de même qu'un énoncé ne contenant que des *OOV* ou *SPR*. Ce dernier, lors de l'analyse en concepts, produit comme résultat une séquence vide qui est interprétée ensuite comme étant un rejet.
Pour chaque corpus de test nous présentons deux évaluations de la *1-best* : la première sans détection de commentaires, les annotations dans la référence sont donc ignorées ; la deuxième avec détection des commentaires, et donc les séquences de mots de la référence annotées comme étant des commentaires sont remplacées par l'étiquette <*COMMENTAIRE*> (elle n'allume aucun concept). Le tableau 4.11 présente l'évaluation sur le corpus **Test_I** et le tableau 4.12 l'évaluation sur le corpus **Test_II**. Les deux tableaux présentent aussi le taux de mots correct ainsi que les détails des erreurs : le taux de faux rejet (FR), de substitution (Sub) et de fausse alarme (FA).

Test_I	IER	C	FR	Sub	FA
sans détection des commentaires	22.0%	86.0%	10.8%	3.3%	7.9%
avec détection des commentaires	20.4%	85.7%	7.6%	6.7%	6.1%

TABLE *4.11* – *Evaluation de la* 1-best *sur le corpus* ***Test_I*** *sans et avec détection des commentaires*

Test_II	IER	C	FR	Sub	FA
sans détection des commentaires	22.7%	92.7%	1.6%	5.7%	15.4%
avec détection des commentaires	24.6%	92.8%	1.6%	5.5%	17.5%

TABLE *4.12* – *Evaluation de la* 1-best *sur le corpus* ***Test_II*** *sans et avec détection des commentaires*

Nous rappelons que sur un corpus de test les deux évaluations, avec et sans détection de commentaires ne peuvent pas être comparées directement du fait que la référence n'est pas la même. En effet, l'*IER* est calculé par rapport au nombre d'énoncés interprétables (la référence produit une interprétation valide). Comme nous l'avons expliqué au 4.1 les séquences mots annotées comme étant des commentaires peuvent contenir des mots qui allument des concepts et qui peuvent donner lieu à une interprétation. Ceci est le cas lorsqu'on ignore ces annotations comme pour l'évaluation sans détection de commentaires. Lorsque ces séquences sont remplacées par l'étiquette <*COMMENTAIRE*>, elles n'allument aucun concept et aucune interprétation n'est alors possible. Le nombre d'énoncés interprétables n'est donc pas le même sur les deux évaluations. Le tableau 4.13 montre le nombre d'énoncés interprétables pour les deux évaluations. On observe bien une différence qui provient du fait que pour 114 énoncés, les séquences de mots annotées comme étant des commentaires dans la référence produisent une interprétation si l'annotation est ignorée.

4.6 Conclusions

Dans ce chapitre nous avons présenté le cadre expérimental des travaux de cette thèse. Nous avons tout d'abord réalisé une analyse des énoncés utilisateur et nous

Test_II	Nombre énoncés interprétables
sans détection des commentaires	4253
avec détection des commentaires	4139

TABLE *4.13* – *Nombre d'énoncés interprétables pour le corpus **Test_II** avec et sans détection de commentaires*

avons mis en évidence plusieurs catégories d'énoncés comme les énoncés non-parole (ne contenant que du bruit), la parole *Hors-Domaine* (les commentaires) et la parole *Dans-le-Domaine*. Nous avons aussi posé les problématiques liées au comportement du système de dialogue en fonction du type d'énoncé traité et nous avons introduit la notion d'énoncé à rejeter. Nous avons ensuite présenté le modèle de langage et le modèle acoustique utilisé par le SRAP. Un sous-modèle de langage pour la détection des commentaires a également été introduit. Une description des données expérimentales (corpus d'apprentissage, de développement et de test) a aussi était donnée.
Une problématique importante abordée dans ce chapitre est liée à l'évaluation au niveau mot des corpus réels. Nous avons ainsi présenté les différents problèmes posés par le traitement des données réelles et nous avons décrit quatre méthodes de normalisation de données qui permettent le calcul du *WER* sur un ensemble homogène. Une évaluation des corpus de test en termes de *WER* a aussi été présenté. Nous n'avons retenu que les deux dernières méthodes pour la suite des évaluations. L'évaluation des corpus de test en termes de *IER*a également été présentée.