Stratégies de décision

Sommaire

Dans une application de dialogue, la compréhension d'un message est réalisée par l'analyse de la transcription issue du module de reconnaissance. En ce qui concerne le service 3000, l'interprétation de cette transcription passe par une succession d'étapes contrôlées par des sources de connaissance imparfaites (les règles d'allumage de concepts, les règles d'interprétation) qui traitent des données qui peuvent être incorrectes (les hypothèses de reconnaissance). Pour cette raison il est important dans ce type d'application de réduire la probabilité d'apparition des erreurs le plus tôt possible dans le déroulement du processus.
Pour des systèmes de dialogue en langage naturel, tel que le service 3000, déployé auprès du grand public, une analyse des données nous a montré qu'il existe différents types de messages, comme les énoncés non-parole, les segments de parole *Hors-Domaine* et la parole *Dans-le-Domaine*. Comme nous l'avons expliqué dans la section 4.1, les segments *Hors-Domaine* peuvent générer une interprétation et orienter le système dans une mauvaise direction pour la suite du dialogue. De plus la génération des CNs pour ces énoncés, mais aussi pour les énoncés non-parole, est très coûteuse en termes de temps de calcul car les graphes correspondant sont très bruités et donc très "volumineux". Le traitement de données réelles implique donc de devoir prendre en compte

et de pouvoir traiter certains types d'énoncés qui, du point de vue du système, sont à rejeter. Pour pallier cela nous proposons de rejeter ces énoncés dès la première passe de reconnaissance. Les CNs sont générés sur les énoncés considérés valides. Concernant le processus d'interprétation, il est, dans un premier temps, appliqué sur la meilleure solution du CN pour tous les CNs générés. Au delà de cette approche séquentielle, nous proposons également une stratégie qui permet l'utilisation de la recherche intégrée sur le graphe de mots et sur les CNs, décrite dans la section 3.3.3, dans le processus d'interprétation.
Le chapitre est organisé en trois parties. Dans la section 7.1, nous présentons les changements apportés au protocole expérimental du fait de la prise en compte des spécificités liées au traitement des données réelles. Une évaluation détaillée du comportement de l'algorithme du *pivot topologique* sur les différents types d'énoncés est aussi présentée. Dans la section 7.2, nous présentons une stratégie de décision à plusieurs niveaux (Minescu et Damnati, 2008) qui vise à rejeter les énoncés non-valides (bruits, commentaires) dès la première passe de reconnaissance et à construire les CNs seulement sur les énoncés valides. L'approche séquentielle (analyse en concepts suivie de l'analyse sémantique) du processus d'interprétation est appliquée sur la *consensus hypothesis* pour tous les CNs générés. Dans la section 7.3, nous présentons une stratégie de décision basée sur une approche intégrée [1] du processus d'interprétation. Cette approche introduit un niveau de décision supplémentaire de rejet des énoncés valides mais qui ne sont pas couverts par le modèle sémantique de l'application (aucune interprétation n'est trouvée).

7.1 Cadre expérimental

Dans les évaluations présentées dans les chapitres précédents, nous n'avions pas pris en compte la variabilité des données réelles utilisées. Par exemple, le sous-modèle de langage de détection des commentaire n'a pas été utilisé par la première passe de reconnaissance. Ainsi, les énoncés *Hors-Domaine* n'ont pas été traités différemment des autres énoncés et, par conséquence, les annotations des commentaires réalisées dans les transcriptions manuelles des énonces ont été ignorées. Les mots composant les commentaires ont été comptabilisés comme tous les autres mots de l'application.
Le tableau 7.1 présente deux énoncés *Hors-Domaine* avec la transcription manuelle annotée. Lorsque aucun traitement particulier n'est appliqué aux énoncés *Hors-Domaine* (on ne détecte pas les commentaires) la référence est constituée des mots présents dans la transcription manuelle (les annotations sont ignorées). En revanche, lorsqu'on souhaite appliquer des traitement spécifiques pour les énoncés *Hors-Domaine* (on détecte les commentaires), dans la référence de l'énoncé on remplace la séquence de mots, annotée comme étant un commentaire dans la transcription manuelle, par un symbole unique <*COMMENTAIRE*>. Afin de pouvoir réaliser une évaluation du *WER* qui soit pertinente dans le contexte du système de dialogue, nous utilisons la méthode de normalisation **Methode 4**, décrite dans la section 4.4.1. Les énoncés ne contenant que des

1. La recherche intégrée sur les graphes de mots et sur les CNs, décrite dans la section 3.3.3, est utilisée pour obtenir une interprétation.

<*COMMENTAIRE*>, des *OOV* ou des *SPR* sont considérés comme étant un rejet et la référence est remplacée par le symbole <*REJET*>. Nous utilisons cette normalisation afin de ne pas compter comme une erreur de reconnaissance une substitution du type <*COMMENTAIRE*> → *OOV*.

	Exemple 1	Exemple 2
Énonce	oh merde oh là oh	eh ben France Telecom
Transcription manuelle annotée	[com :] oh merde oh là oh [:com]	[com :] eh ben France Telecom [:com]
Référence sans prise en compte des commentaires	oh merde oh là oh	eh ben France Telecom
Référence avec prise en compte des commentaires	<*COMMENTAIRE*>	<*COMMENTAIRE*>

TABLE 7.1 – *L'influence de la prise en compte des commentaires sur la référence d'un énoncé* Hors-Domaine *au niveau mot.*

Interprétation	Exemple 1	Exemple 2
Énonce	Rejet	Serv(OffresFT)
Référence sans prise en compte des commentaires	Rejet	Serv(OffresFT)
Référence avec prise en compte des commentaires	Rejet	Rejet

TABLE 7.2 – *L'influence de la prise en compte des commentaires sur l'interprétation de la référence d'un énoncé* Hors-Domaine.

Le changement de la référence au niveau mot à aussi des répercussions au niveau interprétation. Ainsi, comme le montre le tableau 7.2, lorsqu'on souhaite appliquer des traitement spécifiques pour les énoncés *Hors-Domaine* (on détecte les commentaires), les énoncés étiquetés comme étant des commentaires deviennent des énoncés à rejeter. Dans le cas contraire, les mots de la référence peuvent produire une interprétation, comme pour l'exemple 2. Sur le corpus **Test_II** on observe ainsi une augmentation du nombre d'énoncés à rejeter de 2% (soit 114 énoncés) du nombre total d'énoncés dans le corpus. Ce pourcentage représente le nombre d'énoncés couverts par l'analyse sémantique (qui donnent lieu à une interprétation) si aucun traitement spécifique n'est appliqué aux énoncés *Hors-Domaine* (la première passe de reconnaissance ne détecte pas les commentaires). L'exemple 2 dans le tableau 7.2 illustre bien cette situation.

7.1.1 Description du corpus de test par catégorie d'énoncés

Dans la partie 4.1, l'analyse des données réelles issues du service 3000 nous a permis de distinguer trois catégories principales d'énoncés : les énonces non-parole (**C1-Non-Parole**), la parole *Hors-Domaine* (les commentaires, **C2-Hors-Domaine**) et la parole *Dans-le-Domaine* (**C3-Dans-le-Domaine**). Le tableau 7.3 présente une description du corpus de **Test_II** par catégorie, en détaillant le nombre d'énoncés par catégorie ainsi que la taille des graphes correspondants exprimée en nombre moyen de transitions par graphe.

Catégorie	# énoncés	# transitions par graphe
C1-Non-Parole	1333	24000
C2-Hors-Domaine	674	23000
C3-Dans-le-Domaine	4494	14000
Total	6501	17000

TABLE 7.3 – *Description du corpus de test **Test_II** par catégorie*

On observe que pour les deux premières catégories, **C1-Non-Parole** et **C2-Hors-Domaine**, le moteur de reconnaissance a tendance à être sur-génératif. Les graphes générés sont, en moyenne, 70% plus grands que pour la catégorie **C3-Dans-le-Domaine**. Non seulement la génération de ces graphes mais aussi la génération des CNs pour ces énoncés est très coûteuse en terme de temps de calcul car les graphes correspondant sont très bruités et redondants et donc très "volumineux". De ce fait, la génération des graphes de mots n'est pas envisageable en situation réelle pour les énoncés appartenant à ces deux catégories.

7.1.2 Analyse de l'algorithme du *pivot topologique* par catégorie d'énoncés

Le tableau 7.4 [2] montre le détail du *WER* sur chaque catégorie pour la *1-best* et la *consensus hypothesis* extraite à partir de l'algorithme du *pivot topologique*. Le *WER* sur chaque catégorie est calculé comme une contribution sur le *WER* de l'ensemble du corpus. La dernière colonne du tableau représente le nombre d'erreurs (substitutions, insertions et omissions) pour chaque catégorie.
Les performances en terme de *WER* sont globalement dégradées avec les CNs issus de l'algorithme du *pivot topologique*, la dégradation étant plus significative sur **C1-Non-Parole** et **C2-Hors-Domaine**. Malgré le fait que ces deux catégories ne représentent que 30% des énoncés, elle contribuent pour près de la moitié aux erreurs observées avec les CNs. De plus, comme le montre le tableau 7.5, la taille des CNs sur ces deux catégories est deux à trois fois supérieure à celle des CNs de la catégorie **C3-Dans-le-Domaine**. En effet, les graphes "volumineux" et hautement ambigus génèrent à leur tour des CNs bruités, pour lesquels la *consensus hypothesis* est source d'insertions et de substitutions. La génération de ces CNs est aussi trop lente est trop couteuse en terme de ressources mémoire. Ces résultats illustrent l'inadéquation des CNs pour rejeter des entrées non-valides.

2. Le *WER* de l'algorithme du *pivot topologique* présenté dans le tableau 6.2 a été évalué en comptant tous les mots des séquences annotées en commentaires, ce qui n'est plus le cas ici, car ces séquence ont été remplacées par le symbole <*COMMENTAIRE*>. Le nombre de mots dans la référence est donc différent ainsi que la valeur du *WER*, qui était de 49%.

3. Le *WER* calculé pour chaque catégorie est une contribution du *WER* total. On divise le nombre d'erreurs de chaque catégorie par le nombre total de mots de la référence sur l'ensemble du corpus.

		WER [3]	Nombre d'erreurs
C1-Non-Parole	1-best	6.8%	914
1333 én.	*pivot topologique*	11.6%	1549
C2-Hors-Domaine	1-best	10.9%	1452
674 én.	*pivot topologique*	12.8%	1704
C3-Dans-le-Domaine	1-best	24.1%	3236
4494 én.	*pivot topologique*	26.9%	3584
Total	1-best	41.8%	5602
6501 én.	*pivot topologique*	51.3%	6837

TABLE 7.4 – WER *de la* 1-best *et de la* consensus hypothesis *issue de l'algorithme du* pivot topologique.

	# classes / mot de la référence	# mots / classe
C1-Non-Parole	65.6	6.3
C2-Hors-Domaine	51.8	7.0
C3-Dans-le-Domaine	23.0	5.6
Total	34.8	5.9

TABLE 7.5 – *Taille des CNs issus de l'algorithme du* pivot topologique.

7.2 Stratégie de décision basée sur une approche séquentielle

Comme le montre l'analyse effectuée dans la section précédente, nous devons envisager un moyen efficace et rapide qui permette au système de rejeter les énoncés non-valides (les énoncés non-parole et les énoncés *Hors-Domaine*) sans avoir besoin de générer les graphes de mots ni par conséquent les CNs correspondants. La *1-best* présente une précision de 90.2% pour la détection des énoncés non-parole. Elle constitue ainsi un moyen fiable et rapide. De plus, le sous-modèle de détection des commentaires lui permet de détecter de manière efficace les énoncés *Hors-Domaine*. En situation réelle nous commencons la génération du graphe de mots en parallèle avec la *1-best*. En conséquence nous avons décidé d'utiliser la *1-best* afin de construire une stratégie qui vise à rejeter les énoncés non-valides (appartenant à **C1-Non-Parole** et **C2-Hors-Domaine**) et à ne générer les CNs que sur les énoncés contenant de la parole *Dans-le-Domaine* de l'application. Les CNs ne sont ainsi générés que si la première passe ne détecte ni un énoncé non-parole ni un commentaire. Dans le cas contraire, la génération du graphe de mots est arrêtée.

7.2.1 Étapes de la stratégie

La figure 7.1 montre les trois étapes de la stratégie. Pour les deux premières étapes, la décision est prise à partir de la *1-best* et c'est seulement lors de la troisième étape que les graphes de mots et les CNs correspondant sont construits. Le déroulement de la stratégie est le suivant :

Étape 1 : *Est-ce que la* 1-best *contient de la parole ?*

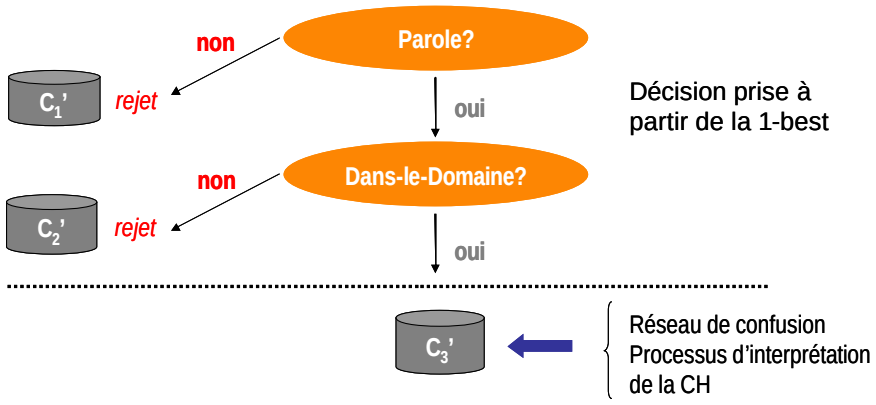

FIGURE 7.1 – Les différentes étapes de la stratégie de construction des CNs

- Si la réponse est non, il s'agit alors d'une détection d'un énoncé non-parole. Étant considéré comme un message non-valide, la stratégie le rejette et le classe dans la catégorie $C_1^{'}$.
- Si la réponse est oui, on passe à l'étape suivante.

Étape 2 : *Est-ce que la* 1-best *contient de la parole* Dans-le-Domaine *?*

- Si la réponse est non, il s'agit alors d'une détection de commentaires. Étant considéré comme un message non-valide, la stratégie le rejette et le classe dans la catégorie $C_2^{'}$.
- Si la réponse est oui, on passe à l'étape suivante.

Étape 3 : Tous les énoncés traités lors de cette étape sont considérés comme faisant partie de la catégorie $C_3^{'}$ ce qui signifie que la *1-best* est supposée contenir de la parole *Dans-le-Domaine*. Pour les énoncés dont la *1-best* ne contient que de la parole *Dans-le-Domaine*, la stratégie génère les réseaux de confusion et extrait la *consensus hypothesis*. Celle-ci est ensuite soumise au processus d'interprétation. [4].

Les catégories $C_1^{'}$, $C_2^{'}$ et $C_3^{'}$ ne sont pas identiques aux catégories **C1-Non-Parole**, **C2-Hors-Domaine** et **C3-Dans-le-Domaine**. Elles ont la même fonction, celle de regrouper les énoncés non-parole, les énoncés *Hors-Domaine* et respectivement les énoncés contenant de la parole *Dans-le-Domaine*, mais elles ne regroupent pas exactement les mêmes énoncés. En effet, les catégories **C1-Non-Parole**, **C2-Hors-Domaine** et **C3-Dans-le-Domaine** ont été calculées en fonction de la référence des énoncés, pour des raisons d'analyse et évaluation du corpus de test, alors que les catégories $C_1^{'}$, $C_2^{'}$ et $C_3^{'}$ sont calculées en fonction de la *1-best*.

4. Un traitement particulier est réservé aux énoncés dont la *1-best* contient des commentaires et de la parole *Dans-le-Domaine*. Comme nous l'avons expliqué dans la section 4.3, le sous-modèle de détection des commentaires n'est pas utilisé dans la génération des graphes de mots car il est difficile d'estimer la probabilité *a posteriori* sur les segments *Hors-Domaine* et donc de générer les CNs correspondant. La *1-best* est alors utilisée comme hypothèse de reconnaissance pour ce type d'énoncé et est ensuite soumise au processus d'interprétation.

7.2.2 Evaluation

Un premier jeu d'expériences présenté dans le tableau 7.6 vise à montrer les performances de la stratégie en terme de *WER*. On observe un gain absolu de 7% en terme de *WER* sur l'algorithme du *pivot topologique* grâce à l'utilisation de la stratégie. Toutefois, le *WER* de la *1-best* reste légèrement meilleur que celui obtenu avec la stratégie, la différence provenant essentiellement des énoncés appartenant à **C3-Dans-le-Domaine**.

WER	Total	**C1-Non-Parole**	**C2-Hors-Domaine**	**C3-Dans-le-Domaine**
1-best	41.8%	6.8%	10.9%	24.1%
pivot topologique	51.3%	11.6%	12.8%	26.9%
Stratégie + Alg. du *pivot topologique*	44.3%	6.6%	10.8%	26.9%

TABLE 7.6 – *Evaluation de la stratégie en terme de* WER.

Le tableau 7.7 montre le détail du *WER*, obtenu sur le trois catégories, de la stratégie utilisant l'algorithme du *pivot topologique*, l'algorithme *multi-niveaux* ou *multi-niveaux sans vides* pour générer les CNs sur les enregistrements du $C_3^{'}$. Le tableau 7.8 montre le détail d'*IER* sur les trois catégories. Nous rappelons que les taux de fausse alarme, de substitutions et de faux rejets sont calculés par rapport au nombre d'énoncés interprétables.

WER	Total	**C1-Non-Parole**	**C2-Hors-Domaine**	**C3-Dans-le-Domaine**
1-best	41.8%	6.8%	10.9%	24.1%
Stratégie + Alg. du *pivot topologique*	44.3%	6.6%	10.8%	26.9%
Stratégie + Alg. *multi-niveaux*	42.3%	6.3%	10.2%	25.8%
Stratégie + Alg. *multi-niveaux sans vides*	38.6%	5.3%	8.7%	24.6%

TABLE 7.7 – *Les performances en terme de* WER *de la stratégie en fonction de l'algorithme de génération des CNs utilisé à l'étape 3.*

L'utilisation de l'algorithme *multi-niveaux* à la place de l'algorithme du *pivot topologique* dans l'Etape 3 de la stratégie permet d'obtenir une diminution du *WER* de 2%, en valeur absolue. L'utilisation de l'algorithme *multi-niveaux sans vides* permet d'améliorer encore plus les performances de la stratégie avec une réduction absolue de 3.2% du *WER* par rapport à la *1-best*. En ne conservant que les mots vides présents initialement dans le *pivot*, on évite d'augmenter de façon trop importante le nombre d'insertions des mots vides. Ceci à d'autant plus d'effet sur les catégories **C1-Non-Parole** et **C2-Hors-Domaine** qui gèrent les graphes les plus bruités. L'amélioration des performances par rapport à la *1-best* sur les catégories **C1-Non-Parole** et **C2-Hors-Domaine** est due aux énoncés, appartenant à ces deux catégories, classés à tort par la *1-best* dans la catégorie $C_3^{'}$ et pour lesquels les CNs sont générés (30% pour **C1-Non-Parole** et 80% pour **C2-Hors-Domaine**, cf. tableau 7.9). En effet, sur ces énoncés, on observe une baisse relative

		C1-Non-Parole	C2-Hors-Domaine	C3-Dans-le-Domaine		
	IER	FA	FA	FA	Sub	FR
1-best	24.6%	4.6%	8.8%	4.1%	5.5%	1.6%
Stratégie + Alg. du *pivot topologique*	22.0%	2.4%	6.2%	3.2%	6.0%	4.2%
Stratégie + Alg *multi-niveaux*	20.1%	2.2%	6.3%	3.1%	5.1%	3.4%
Stratégie + Alg. *multi-niveaux sans vides*	19.9%	2.1%	6.2%	3.1%	5.1%	3.4%

TABLE 7.8 – *Les performances en terme d'*IER *de la stratégie en fonction de l'algorithme de génération des CNs utilisé à l'étape 3.*

de 20% des insertions et de 15% des substitutions dans la *consensus hypothesis* par rapport à la *1-best*. La légère dégradation du *WER* sur **C3-Dans-le-Domaine** est due à une augmentation relative de 25% du nombre d'omissions dans la *consensus hypothesis*. Sur cette catégorie, la baisse relative du nombre d'insertions et de substitutions et de 12% en moyenne. La diminution du nombre d'insertions et de substitutions sur les trois catégories explique la diminution du taux de fausse alarme. L'augmentation du taux de faux rejet est due à l'augmentation du nombre d'omissions sur **C3-Dans-le-Domaine**.

Total 6501 én.	C_1' 994 én.	C_2' 138 én.	C_3' 5369 én.	Rappel
C1-Non-Parole 1333 én.	897	47	389	67.3%
C2-Hors-Domaine 674 én.	76	54	544	8.0%
C3-Dans-le-Domaine 4494 én.	21	37	4436	98.7%
Précision	90.2%	39.1%	82.6%	

TABLE 7.9 – *Précision et Rappel pour la classification des énonces par catégorie*

Le tableau 7.9 montre le détail de la classification des énoncés dans chaque catégorie. Pour chacune des trois catégories nous calculons le rappel et la précision en utilisant les formules suivante :

$$\text{Rappel} = \frac{\text{\# énoncés bien détectés}}{\text{\# total d'énoncés dans une classe}} \tag{7.1}$$

$$\text{Précision} = \frac{\text{\# énoncés bien détectés}}{\text{\# total d'énoncés détectés comme appartenant à une classe}} \tag{7.2}$$

Suite à l'utilisation de la stratégie, 17% du nombre total d'énoncés dans le corpus de test ont été rejetés par la *1-best* dans les deux premières étapes de la stratégie, par rapport à 30% des énoncés à rejeter. La précision du rejet de la stratégie, due à l'utilisation de la *1-best*, sur les deux premières catégories est de 95%. Sur la catégorie **C3-Dans-le-Domaine** on obtient un rappel de 98.7%, ce qui signifie que la quasi totalité des énoncés

Dans-le-Domaine ont été correctement traités par la stratégie, en construisant les CNs correspondants.

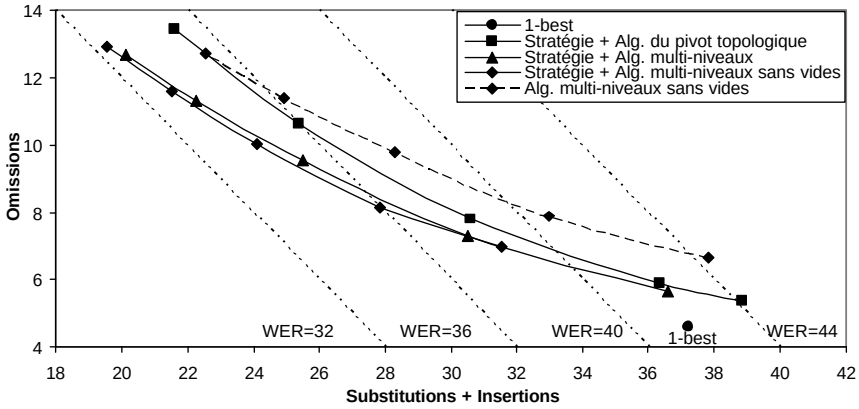

FIGURE 7.2 – Evaluation du *WER* de la stratégie en fonction de la variation du seuil sur la mesure de confiance pour la *consensus hypothesis*

Comme nous l'avons mentionné, dans la troisième étape de la stratégie, la *consensus hypothesis* peut être filtrée en fonction d'un seuil sur la mesure de confiance avant d'appliquer le processus d'interprétation. La figure 7.2 montre les performances de la stratégie en fonction du seuil sur la mesure de confiance qui permet de filtrer les mots de la *consensus hypothesis* extraite à partir des CNs générés à l'étape trois. Les courbes ont été construites en variant le seuil de 0 à 0.4, chaque point sur la courbe correspond à une valeur en partant de la droite vers la gauche. Dans cette figure, nous avons représenté également les performances obtenues en construisant les CNs à l'aide de l'algorithme *multi-niveaux sans vides* sur l'ensemble du corpus de test [5]. On observe des performances équivalentes en terme de *WER* pour la stratégie suite à l'utilisation de l'algorithme *multi-niveaux* ou de l'algorithme *multi-niveaux sans vides* à l'étape trois et l'utilisation d'un seuil sur la mesure de confiance afin de filtrer la *consensus hypothesis* correspondante. Par ailleurs, on observe que la stratégie permet une amélioration importante des performances par rapport à l'utilisation de l'algorithme *multi-niveaux sans vides* sur l'ensemble de corpus. Ainsi, pour un taux d'omission constant on observe une baisse de près de 4% du nombre d'insertions et de substitutions.
Un deuxième jeu d'expériences vise à montrer les performances de la stratégie en termes d'*IER* et le déplacement du point de fonctionnement du système en fonction du seuil sur la mesure de confiance. Les courbes de la figure 7.3 ont été construites de la même manière que pour la figure précédente, en variant le seuil sur la mesure de confiance. Le filtrage des mots de la *consensus hypothesis* en fonction du seuil sur la mesure de confiance est réalisé avant le processus d'interprétation. On observe une amélioration de l'*IER* en utilisant la stratégie avec l'algorithme du *pivot topologique*. Le point de fonctionnement est déplacé vers la gauche dû principalement à la réduction importante du nombre de FA. L'utilisation des algorithmes *multi-niveaux* et *multi-niveaux sans vides*

5. Cette courbe est différente de celle présentée dans la figure 6.2 car la référence est différente du fait de la prise en compte de la détection des commentaires.

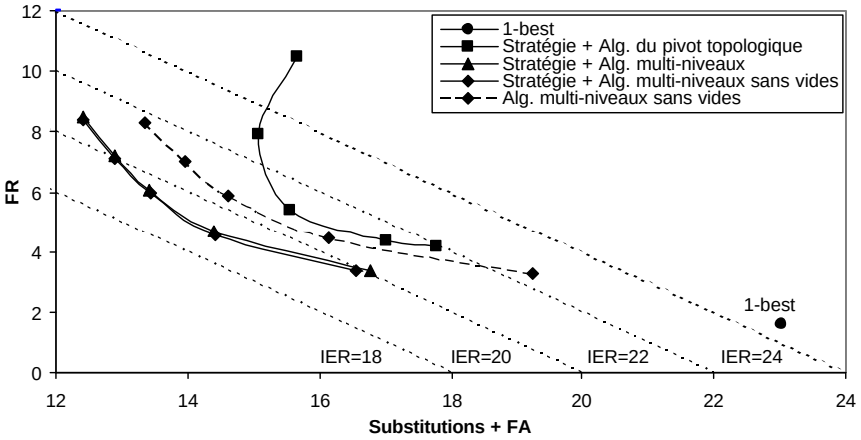

FIGURE 7.3 – La variation du point de fonctionnement de la stratégie en fonction du seuil sur la mesure du confiance

permet d'améliorer d'avantage l'*IER*. Le point de fonctionnement est déplacé encore plus vers la gauche en réduisant d'avantage le nombre de FA et de substitutions (cf. tableau 7.8). On observe également que la stratégie utilisant l'algorithme *multi-niveaux sans vides* dans la troisième étape permet une réduction importante de l'*IER* par rapport à l'utilisation de l'algorithme *multi-niveaux sans vides* sur l'ensemble du corpus. Ainsi, pour un taux de FR constant, on observe une réduction moyenne de 2% du nombre de substitutions et FA.

7.3 Stratégie de décision basée sur une approche intégrée

La stratégie présentée dans la section précédente permet non seulement de ne construire des CNs que sur les énoncés valides mais aussi d'améliorer les performances du système au niveau interprétation. Pour ce faire elle utilise la meilleure solution de la première passe de reconnaissance, capable de détecter et rejeter les énoncés non-parole et la parole *Hors-Domaine*, et ensuite la meilleure solution des CNs qui fournit une interprétation sur les énoncés valides. Toutefois, le module de compréhension travaille directement sur une séquence de mots et ne bénéficie pas de la richesse de l'espace de recherche correspondant (CN ou graphe de mots). De plus, on a montré que l'oracle calculé sur ces structures améliore de manière importante les performances en terme de *WER* (voir la section 6.3.2). On peut donc envisager qu'en retardant la prise de décision (ne pas calculer la meilleure séquence de mots avant le calcul de l'interprétation) en calculant une interprétation sur l'ensemble de l'espace des solutions on puisse améliorer les performances du système de dialogue en termes d'*IER*.
Le projet européen LUNA (voir 3.3.3 pour la présentation du projet) propose une architecture pour un système de dialogue qui se base sur la recherche intégrée pour calculer une hypothèse d'interprétation fournie ensuite au gestionnaire du dialogue, avec d'autres informations supplémentaires comme le contexte du dialogue, pour détermi-

ner la meilleure réponse du système à la requête formulée par l'utilisateur. Comme nous l'avons présentée dans la section 3.3.3, la recherche intégrée utilise des graphes de mots (ou autres structures similaires comme les CNs) afin d'obtenir directement une interprétation sans passer par le traitement d'une séquence de mots. La prise de décision de la meilleure hypothèse ne se fait plus au niveau mot mais est retardée au niveau interprétation. Pour ce faire, le graphe de mots est transformé en un graphe de concepts et ensuite une interprétation est obtenue. Les différentes transformations sont effectuées à travers des compositions avec des FSM représentant les règles d'allumage de concepts et les règles d'interprétation.
Dans cette partie nous utilisons la recherche intégrée sur les graphes de mots et sur les CNs construits pour les énoncés valides afin d'améliorer les performances du système. La stratégie présentée dans cette section (Minescu et al., 2007) rejette les énoncés non-valides de la même manière que la stratégie présentée dans la section 7.2 et n'utilise la recherche intégrée que sur les énoncés valides. Nous montrons également qu'une étape supplémentaire est nécessaire pour le rejet des énoncés valides qui ne sont pas couverts par l'analyse sémantique avant de réaliser une recherche intégrée sur les graphes de mots ou sur les CNs correspondants aux énoncés restants.

7.3.1 Analyse du processus d'interprétation

Une analyse plus approfondie de la catégorie **C3-Dans-le-Domaine** montre qu'une partie des énoncés en faisant partie sont des énoncés à rejeter car ils ne sont pas couverts par l'analyse sémantique. En effet, malgré le fait que l'énoncé ne contienne que de la parole *Dans-le-Domaine*, le système ne trouve pas une interprétation valide pour la séquence de mots prononcée. A la différence de la stratégie présentée au 7.2, où la *consensus hypothesis* extraite des CNs était utilisée pour obtenir une interprétation sur les énoncés valides, l'utilisation de la recherche intégrée sur des graphes de mots (ou des CNs) pour trouver un interprétation peut augmenter considérablement le nombre de FA sur les énoncés valides mais non couverts par l'analyse sémantique. Il convient donc de développer une stratégie capable de rejeter tous les énoncés non-valides, que ce soit des détections non-parole, de la parole *Hors-Domaine* mais aussi les énoncés valides non couverts par l'analyse sémantique.

Catégorie	# énoncés	# transitions par graphe
C1-Non-Parole	1333	24000
C2-Hors-Domaine	674	23000
C3-Dans-le-Domaine-sans-interprétation	355	17000
C3-Dans-le-Domaine-avec-interprétation	4139	13000
Total	6501	17000

TABLE *7.10* – *Description du corpus de test **Test_II** par catégorie*

Pour développer cette stratégie nous faisons la différence entre les énoncés de la catégorie **C3-Dans-le-Domaine** avec et sans interprétation afin de les traiter différemment. Le tableau 7.10 présente la nouvelle découpe par catégorie du corpus **Test_II**.

Les deux premières catégories restent inchangées, alors que la catégorie **C3-Dans-le-Domaine** est divisée en deux : la catégorie **C3-Dans-le-Domaine-sans-interprétation** qui contient les énoncés *Dans-le-Domaine* pour lesquels le processus d'interprétation ne produit aucune interprétation (ils ne sont pas couvert par l'analyse sémantique), qui sont donc à rejeter, et la catégorie **C3-Dans-le-Domaine-avec-interprétation** qui contient les énoncés *Dans-le-Domaine* ayant une interprétation valide. On observe que la taille des graphes de mots des énoncés appartenant à **C3-Dans-le-Domaine-sans-interprétation** est 25% plus grande que celle des énoncés valides couverts par l'analyse sémantique.
Les résultats en terme d'*IER* de la stratégie de construction des CNs, présentés dans le tableau 7.8, montrent que l'utilisation de l'algorithme *multi-niveaux sans vides* pour obtenir une interprétation sur la catégorie $C_3^{'}$ donne de meilleurs résultats que la *1-best* ou l'algorithme du *pivot topologique*, notamment grâce à la réduction du nombre de FA. Cette réduction du nombre de FA correspond en effet à un meilleur rejet par rapport à la *1-best* des énoncés sans interprétation valide appartenant à la catégorie **C3-Dans-le-Domaine** (les énoncés non couverts par l'analyse sémantique). De plus, l'utilisation des mesures de confiance pour filtrer la *consensus hypothesis* avant le processus d'interprétation permet de diminuer davantage le nombre de FA, comme montré dans la figure 7.3. Les performances en terme d'*IER* de l'algorithme *multi-niveaux* sont quasiment équivalentes mais les CNs générés sont plus grands et leur génération nécessite un temps de calcul plus élevé. La *consensus hypothesis* des CNs générés à l'aide de l'algorithme *multi-niveaux sans vides* constitue donc le meilleur choix pour rejeter les énoncés appartenant à la catégorie **C3-Dans-le-Domaine-sans-interprétation**.

7.3.2 Étapes de la stratégie

Nous avons identifié trois catégories d'énoncés à rejeter avant de pouvoir utiliser la recherche intégrée sur les énoncés restant. La stratégie utilise successivement différentes solutions pour rejeter les énoncés avant de calculer une interprétation sur les graphes de mots ou sur les CNs.
La figure 7.4 montre les quatre étapes de la stratégie et les solutions utilisées pour la prise de décision. On observe que le rejet dans les deux premières étapes est réalisé à l'aide de la *1-best* de la même manière que pour la stratégie décrite au 7.2. Le déroulement de la stratégie pour les deux dernières étapes est le suivant :

Étape 3 : Pour les énoncés dont la *1-best* ne contient que de la parole *Dans-le-Domaine*, la stratégie génère le graphe de mots et le CN correspondant utilisant l'algorithme *multi-niveaux sans vides* et extrait la *consensus hypothesis*. Celle-ci est ensuite soumise au processus d'interprétation. [6].

6. Un traitement particulier est réservé aux énoncés dont la *1-best* contient des commentaires et de la parole *Dans-le-Domaine*. Comme nous l'avons expliqué dans la section 4.3, le sous-modèle de détection des commentaires n'est pas utilisé dans la génération des graphes de mots car il est difficile d'estimer la probabilité *a posteriori* sur les segments *Hors-Domaine* et donc de générer les CNs correspondant. La *1-best* est alors utilisée comme hypothèse de reconnaissance pour ce type d'énoncé et est ensuite soumise au processus d'interprétation.

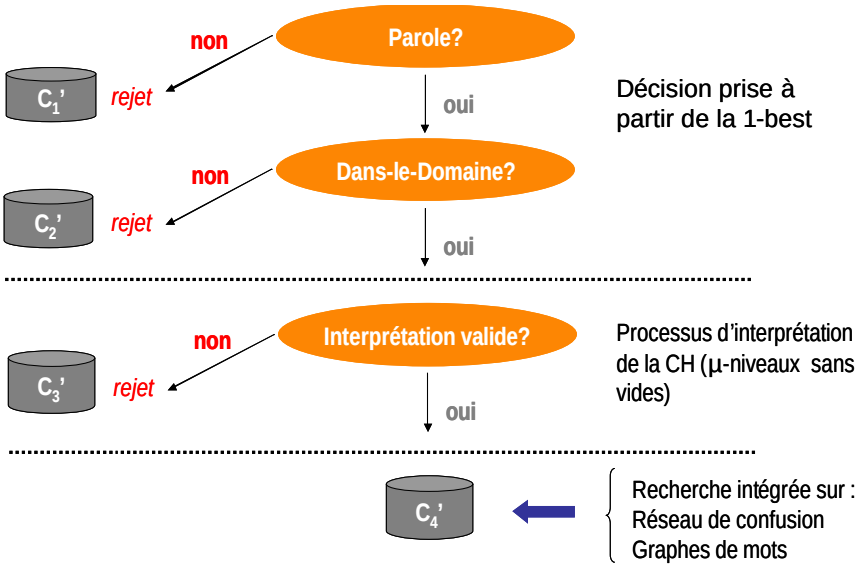

FIGURE 7.4 – Description des différentes étapes de la stratégie

Est-ce que le processus d'interprétation de la consensus hypothesis *a fourni une interprétation valide ?*

- Si non, l'énoncé est rejeté. Étant considéré comme un message *Dans-le-Domaine* sans interprétation valide, la stratégie le rejette et le classe dans la catégorie $C_3^{'}$.
- Si oui, la stratégie passe à l'étape suivante.

Étape 4 : Tous les énoncés traité lors de cette étape sont considérés comme faisant partie de la catégorie $C_4^{'}$. Il s'agit des énoncés *Dans-le-Domaine* ayant une interprétation valide. La stratégie calcule alors, à travers la recherche intégrée, la meilleure solution d'interprétation. Le calcul peut se faire sur le graphe de mots ou sur le CN.

Comme nous pouvons l'observer, pour les deux premières étapes, le rejet est effectué au niveau mot à l'aide de la *1-best*, alors que pour l'étape 3, le rejet est effectué au niveau interprétation à l'aide de la *consensus hypothesis* extraite des CNs générés avec l'algorithme *multi-niveaux sans vides*. Les graphes de mots et les CNs sont générés pour tous les énoncés traités à l'étape 3.

7.3.3 Evaluation

Avant d'évaluer les performances de la stratégie proposée, un premier jeu d'expériences vise à montrer les performances de chacun des algorithmes appliqué sur l'ensemble du corpus sans aucune stratégie. Le tableau 7.11 compare les solutions de six algorithmes en détaillant les erreurs au niveau interprétation sur chacune des quatre catégories. Étant donné que les trois premières catégories ne contiennent que des énoncés à rejeter, on ne peut avoir que des FA. En revanche, la dernière catégorie ne contient que des énoncés ayant une interprétation valide et, en conséquence, on ne peut avoir

que des Sub ou FR. La première colonne correspond à l'utilisation de la meilleure solution de la première passe de reconnaissance. Graphe *1-best* est l'interprétation obtenue sur le meilleur chemin du graphe de mots calculée au sens des probabilités *a posteriori* des transitions. Les deux colonnes suivantes correspondent à l'interprétation obtenue sur la *consensus hypothesis* des deux algorithmes. *Décodage Graphes* et *Décodage CN* correspondent aux interprétations obtenues en effectuant une recherche intégrée respectivement sur les graphes de mots et les CNs. Nous avons utilisé les CNs générés avec l'algorithme *multi-niveaux sans vides*.

	1-best	Graphe *1-best*	*pivot topologique*	*multi-niveaux sans vides*	Décodage CN	Décodage Graphes
FA sur **C1-Non-Parole**	4.6%	6.5%	4.7%	**4.5%**	20.3%	23.0%
FA sur **C2-Hors-Domaine**	8.8%	8.1%	6.6%	**6.5%**	14.5%	13.7%
FA sur **C3-Dans-le-Domaine sans-interp**	4.1%	3.0%	3.2%	**3.1%**	7.2%	6.5%
Substitutions sur **C3-Dans-le-Domaine avec-interp**	5.5%	6.1%	6.0%	5.0%	7.9%	5.9%
FR sur **C3-Dans-le-Domaine avec-interp**	1.6%	2.5%	4.0%	3.3%	0.3%	0.4%
Total *IER*	24.6%	26.3%	24.6%	22.5%	50.2%	49.5%

TABLE *7.11* – *Détail des erreurs sur l'ensemble de corpus pour chaque algorithme.*

L'utilisation d'un espace d'hypothèses plus grand, comme les graphes de mots, améliore les performances (en termes de FR et substitutions) sur **C3-Dans-le-Domaine-avec-interprétation** pour *Décodage Graphes* (de 7.1% à 6.3%) par rapport à la *1-best*, mais augmente en même temps le nombre de FA (jusqu'à 23% sur **C1-Non-Parole**). De plus, le temps de calcul pour *Décodage Graphes* est nettement plus élevé que pour la *1-best*. En revanche, la *consensus hypothesis* obtenue avec l'algorithme *multi-niveaux sans vides* donne les meilleurs résultats en ce qui concerne la réduction de nombre de FA sur les trois premières catégories. On observe également que les performances sur **C2-Hors-Domaine** et **C3-Dans-le-Domaine-sans-interprétation** de la Graphe *1-best* sont meilleures que celles de la première passe de reconnaissance mais le taux de substitution et FR sur **C3-Dans-le-Domaine-avec-interprétation** est plus élevé. De plus, l'obtention de Graphe *1-best* est beaucoup plus coûteuse en temps de calcul , surtout sur les graphes bruités.
Comme nous l'avons mentionné, à l'étape trois de la stratégie, le rejet des énoncés est réalisé au niveau interprétation, à l'aide de la *consensus hypothesis*, obtenue avec l'algorithme *multi-niveaux sans vides*, qui peut être ou non filtrée en fonction d'un seuil sur la mesure de confiance avant le processus d'interprétation. Nous avons choisi de filtrer la *consensus hypothesis* avant le processus d'interprétation avec une valeur du seuil égale à 0.1. En effet, les performances en termes d'*IER* de la stratégie basée sur une approche

séquentielle utilisant l'algorithme *multi-niveaux sans vides* à l'étape 3, présentées dans la figure 7.3, montrent une réduction importante du taux de FA (1.7%) pour une valeur du seuil de 0.1. Cette réduction est réalisée sur la catégorie **C3-Dans-le-Domaine** du fait du filtrage de la *consensus hypothesis* avant le processus d'interprétation. Une valeur plus grande du seuil permet certes de diminuer davantage le taux de FA, mais avec une augmentation importante du taux de FR.
Le tableau 7.12 compare les performances de la stratégie utilisant, à l'étape quatre, la recherche intégrée sur les CNs, que nous appelons *Strat1(CN)*, et celles de la stratégie utilisant la recherche intégrée sur les graphes de mots, que nous appelons *Strat2(Graphe)*, aux performances de la meilleure solution de la première passe de reconnaissance, la *1-best*. La *1-best* est utilisée sur l'ensemble du corpus tant pour le rejet des énoncés que pour la recherche de la meilleure interprétation.

total	*1-best*	Strat 1 (CN)	Strat 2 (Graphe)
FA	17.5%	9.6%	9.6%
Sub	5.5%	5.8%	4.2%
FR	1.6%	4.6%	4.6%
IER	24.6%	20%	18.4%

TABLE 7.12 – *Evaluation des deux stratégies proposées*

Une amélioration importante des performances est obtenue grâce à l'utilisation de *Strat1(CN)*. La stratégie permet de réduire le taux de FA de 7.9% en valeur absolue avec une augmentation de seulement 3% du taux de FR. L'utilisation d'un espace d'hypothèses plus grand dans *Strat2(Graphe)* pour les énoncés de la catégorie $C_4^{'}$ permet de diminuer davantage le taux de substitution. L'augmentation du taux de FR dans les deux cas est due aux énoncés de la catégorie **C3-Dans-le-Domaine-avec-interprétation** mal classifiés par la stratégie dans les catégories $C_1^{'}$, $C_2^{'}$ et $C_3^{'}$ (cf. au tableau 7.13).

Total 6501 én.	$C_1^{'}$ 994 én.	$C_2^{'}$ 138 én.	$C_3^{'}$ 1021 én.	$C_4^{'}$ 4348 én.	Rappel
C1-Non-Parole 1333 én.	897	47	325	64	67.3%
C2-Hors-Domaine 674 én.	76	54	327	217	8.0%
C3-Dans-le-Domaine-sans-interprétation 355 én.	14	13	208	120	58.6%
C3-Dans-le-Domaine-avec-interprétation 4139 én.	7	24	161	3947	95.4%
Précision	90.2%	39.1%	20.4%	90.8%	

TABLE 7.13 – *Précision et Rappel pour la classification des énonces par catégorie*

Le tableau 7.13 monte le détail de la classification des énoncés dans chaque catégorie. Suite à l'utilisation de la stratégie, 33% du nombre total d'énoncés ont été rejetés dans les trois premières étapes, par rapport à 36% des énoncés à rejeter. Sur les deux premières catégories, la précision du rejet due à l'utilisation de la *1-best* est de 95%.

Sur la catégorie $C_3^{'}$, l'utilisation de la *consensus hypothesis* permet d'obtenir une précision de rejet de 20.4%. On observe également que 32.5% des énoncés non-valides qui auraient dû être rejetés par la *1-best* sont rejetés par la *consensus hypothesis* car ils sont classifiés dans la catégorie $C_3^{'}$. Cette classification explique en partie la précision peu élevée sur cette catégorie malgré un rappel de 58,6% sur **C3-Dans-le-Domaine-sans-interprétation**. Sur l'ensemble de la stratégie, la précision du rejet est de 91%. Le rappel sur la catégorie **C3-Dans-le-Domaine-avec-interprétation** est de 95.4% ce qui signifie que la majorité des énoncés *Dans-le-Domaine* et couverts par l'analyse sémantique on été correctement traité par la stratégie en réalisant une recherche intégrée afin de trouver l'interprétation.

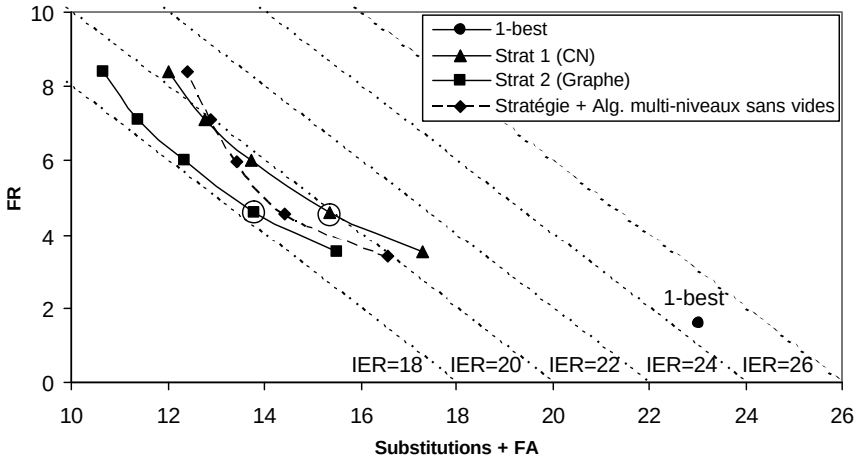

FIGURE 7.5 – Evaluation de la stratégie en fonction de rejet des énoncés à l'étape 3

Dans le tableau 7.12, les deux stratégies présentées utilisent, à l'étape trois, la *consensus hypothesis* obtenue avec l'algorithme *multi-niveaux sans vides* et un filtrage avec une valeur de seuil de 0.1 avant le processus d'interprétation . La figure 7.5 présente la variation du point de fonctionnement des deux stratégies en fonction du seuil sur la mesure de confiance qui est utilisé pour filtrer la *consensus hypothesis* à l'étape 3. Les courbes ont été construites en variant la valeur du seuil de 0 à 0.4, avec le point le plus à droite qui correspond à l'utilisation de la *consensus hypothesis* sans filtrage. Les deux points entourés sur les courbes correspondent respectivement au détail des erreurs dans le tableau 7.12. On observe que le choix d'une valeur du seuil égale à 0.1 est un bon compromis entre l'amélioration du nombre de FA par rapport à la dégradation du nombre de FR. La troisième courbe correspond à la stratégie présentée dans la section 7.2 utilisant l'algorithme *multi-niveaux sans vides* pour construire les CNs dans l'étape trois. On observe ainsi que pour *Strat2(Graphe)*, l'utilisation d'un espace de recherche plus riche (les graphes de mots) combiné à une stratégie de rejet adaptée, permet d'obtenir de meilleures performances par rapport à l'utilisation, sur l'ensemble des énoncés valides, de la *consensus hypothesis* extraite des CNs construits avec l'algorithme *multi-niveaux*. On observe dans la figure 7.5, que *Strat1(CN)* ne permet pas d'améliorer les performances de la stratégie proposée dans la section 7.2 ce qui montre que les CNs ne sont pas adaptés pour réaliser une recherche intégrée de la solution d'interprétation. En re-

vanche, les CNs s'avèrent un moyen très efficace de sélectionner les énoncés pertinents à travers l'utilisation de la *consensus hypothesis*.

7.4 Conclusions

Ce chapitre a introduit, dans un premier temps, un nouveau cadre expérimental défini par la détection des commentaires qui implique une prise en compte, dans la référence, des annotations indiquant les séquences de mots qui représentent des commentaires. Ces séquences sont remplacées par l'étiquette <*COMMENTAIRE*>. Nous avons ensuite réalisée une analyse détaillée des performances de la *1-best* et de l'algorithme du *pivot topologique* sur les trois types d'énoncés (détections non-parole, parole *Hors-Domaine* et parole *Dans-le-Domaine*) qui ont été identifiés dans le corpus de test. Cette analyse nous a permis de montrer que l'utilisation des CNs n'est pas adaptée pour la détection et le rejet des énoncés non-valides. De plus, la génération des CNs sur ces énoncés est très coûteuse en temps de calcul ce qui n'est pas compatible avec les contraintes temps-réel d'un système du dialogue. Ceci nous a amené à proposer une stratégie qui utilise la *1-best* de la première passe de reconnaissance pour rejeter les énoncés non-valides dès la première passe, et ceci avec une précision de 95%. Les CNs sont ensuite générés seulement pour les énoncés valides. Le processus d'interprétation est effectué sur la *consensus hypothesis* issue des CNs, qui peut, ou non, être filtrée en fonction d'un seuil sur la mesure de confiance. Cette stratégie permet d'améliorer le *WER* et l'*IER* par rapport à la *1-best*.
Nous avons montré par ailleurs que le recours à la recherche intégrée pour rechercher la meilleure interprétation sur le graphe de mots ou le CN ne s'avère pas pertinente dans tous les cas. En effet, il subsiste une partie des énoncés valides qui ne sont pas couverts par l'analyse sémantique et qui ne peuvent pas être interprétés par le système. L'utilisation de la recherche intégrée sur les graphes de mots ou sur les CNs correspondant peut donner lieu à des erreurs d'interprétation (notamment des fausses alarmes). Nous avons proposé une stratégie adaptée qui permet de rejeter également ce type d'énoncé en plus des énoncés non-valides grâce à un enchaînement d'étapes de rejet utilisant différentes solutions. Ainsi, pour le rejet des énoncés non-valides la stratégie utilise la meilleure solution de la première passe de reconnaissance. Ensuite, les CNs sont générés et les énoncés pour lesquels la *consensus hypothesis* ne peut pas être interprétée sont rejetés. La recherche intégrée est ensuite effectuée pour les énoncés restants. Les CNs s'avèrent un moyen très efficace de sélectionner les énoncés pertinents en rejetant correctement 58% des énoncés valides non couverts par l'analyse sémantique. Les résultats montrent une amélioration des performances au niveau interprétation par rapport à la stratégie décrite précédemment qui n'utilise que la *consensus hypothesis* pour obtenir une interprétation sur les énoncés valides.